

Real Statistics: A Radical Approach

authentic-dua.com

اللَّهُمَّ انْفَعْنِي بِمَا عَلِمْتَنِي وَ عَلِمْنِي مَا يَنْفَعُنِي وَ زِدْنِي عِلْمًا

**Allaahum-manfa'nee bimā 'allamtanee wa
'allimnee mā yanfa'unee wa zidnee 'ilmā**

O Allah, benefit me with what You have taught me, and teach me that which benefit me, and increase me in knowledge.

Ibn Majah 1/92, see Sahih Ibn Majah 1/47

Translation Source: From the Book "Supplications & Treatment with Ruqyah",
Dr. Sa'eed bin Ali Al-Qahtani, Dar-us-Salam Publication

AbdurRahman.Org

By

Asad Zaman

Preface:

In her book entitled “The Making of the Modern University: Intellectual Transformation and the Marginalization of Morality”, Julie Reuben describes a momentous revolution in the Western intellectual tradition. This was a transition from a view of all knowledge united within the framework of religion to a positivist view, which drew a sharp distinction between values and facts. A detailed quote brings out key elements of this transition:

In 1884 Harvard officials ... had inherited a world view that strongly associated truth and religion. The term truth encompassed all "correct" knowledge; religious doctrines, common-sense beliefs, and scientific theories were all judged by the same cognitive standards. Religious truth was the most important and valuable form of knowledge because it gave meaning to mundane knowledge. Religion transformed abstract knowledge into "moral" truths—truths that guided individuals' daily actions and explained their ultimate destiny.

The Harvard officials' views about truth represented the beliefs of most educated Americans at that time. In the late nineteenth century intellectuals assumed that truth had spiritual, moral, and cognitive dimensions. By 1930, however, intellectuals had abandoned this broad conception of truth. They embraced, instead, a view of knowledge that drew a sharp distinction between "facts" and "values." They associated cognitive truth with empirically verified knowledge and maintained that by this standard, moral values could not be validated as "true." In the nomenclature of the twentieth century, only "science" constituted true knowledge. Moral and spiritual values could be "true" in an emotional or nonliteral sense, but not in terms of cognitively verifiable knowledge. The term truth no longer comfortably encompassed factual knowledge and moral values.

As has been detailed by Julie Reuben, and many other authors, it was the emergence of the philosophy of logical positivism that created this schism: science was classified as the sole source of knowledge, while all other types of human knowledge were reduced in value to meaningless noise. The logical positivists argued that “... *it is impossible to find a criterion for determining the validity of ethical judgements ... because they have no objective validity whatsoever... They are pure expressions of feeling and as such do not come under the category of truth and falsehood.*” As a philosophy, logical positivism had a spectacular crash in mid 20th Century. Unfortunately, the schism in the foundations of knowledge created by it was never healed. Even though the philosophy has been rejected, the methodologies of the social sciences based upon it continue to dominate the universities, and shape the minds of the seekers of knowledge.

Islam strongly supports the idea of unity of knowledge. The value of knowledge, and the honor accorded to seekers of knowledge, is unique to Islamic teachings. More details on this issue are provided in the first chapter. For this preface, we note that it is a unique feature of Islam that human beings are honored because of knowledge. From the beginning, Allah T'aala gave knowledge to Adam AS, and then the angels were ordered to prostrate. In the first revelation to our Prophet Mohammed SAW, Allah Táala introduces Himself as a TEACHER, who will give to mankind a knowledge which they do not have. This knowledge transformed the ignorant and backwards early Muslims into the leaders of the world.

The main thesis of this textbook is that the knowledge revealed to mankind is still just as revolutionary as it was 1440 years ago. However, the emergence of this split in the body of knowledge, the elevation of “science” and the rejection of religion, has led to an unwarranted and unjustified fragmentation of knowledge. Knowledge is unified by purpose. When we make our goal the acquisition of knowledge for the purpose of recognition and glorification of God, and for providing service to the creation of God, this changes both the substance of knowledge and the method of approach to study. This also implies that all human knowledge is built on moral foundations. The idea of “secular knowledge” is an oxymoron. No aspect of human knowledge can be separated from the fundamental knowledge of the existence of the ONE God, which unites all of the existent. These central ideas about the nature of knowledge will be illustrated by this course on “statistics”. It is the use of these Islamic methodological perspectives which has led to a radically different approach to the subjects, which rejects and replaces the approach initiated by Sir Ronald Fisher a century ago.

In this preface, we mention some of the novel ideas which form part of this new approach to the discipline of statistics. It should be emphasized that this is not at all just conventional statistics with a sprinkling of Quran and Hadeeth thrown in for flavor. Sir Ronald Fisher is known as the father of statistics. He created the approach to statistics currently in use in the discipline all over the world. This approach is based on making certain simplifying assumptions about the data which permit handling large data sets with minimal computational capabilities. We reject this approach in toto, for two reasons. One is that computational capabilities today allow us to avoid making unjustified assumptions. The second is deeper: the ability to make unjustifiable assumptions as part of the standard statistical methodology allows one to pursue agendas and hide them within the assumptions. That allows one to easily “lie with statistics”. This capability can be, and has been, used to dramatically misrepresent ground realities, and paint the outcomes of enormously unjust and exploitative economic systems as remarkable achievements. This accounts for the famous saying that “there are lies, damned lies, and statistics”, Our hope in developing this alternative approach is to keep closer track of the relationships between the statistics and the ground realities they represent, so that the Islamic imperatives of telling the truth are better served by this discipline.

This book is dedicated to: My wife, and the love of my life: Iffat Fatima Zaman.

Table of Contents

| | |
|--|----|
| Preface: | ii |
| Chapter 0: Introduction | 1 |
| A: Preliminaries | 1 |
| B: The Puzzles Posed by Human History | 2 |
| C: The Central Importance of Knowledge..... | 3 |
| D: Rebuilding Knowledge on Islamic Foundations..... | 4 |
| E: Methodology and Approach..... | 7 |
| Chapter 1: All Human Knowledge is Built on Moral Foundations | 11 |
| 1A: The Illusion of Objectivity..... | 11 |
| 1B: Purpose: The Heart of An Islamic Approach..... | 14 |
| 1C: Knowledge: Islam vs. West..... | 17 |
| 1D: Islamic Pedagogy: Engaging the Heart..... | 20 |
| 1E An Islamic Methodology and Epistemology..... | 22 |
| Chapter 2: Comparing Numbers | 25 |
| 2A Comparisons Driven by Purpose..... | 26 |
| 2B Arbitrariness of Rankings..... | 29 |
| 2C What Do College Rankings Measure? | 31 |
| 2D Goodhart's Law..... | 34 |
| 2E How Rankings Hide Values | 37 |
| 2F The Why of Global Corruption Rankings | 39 |
| Chapter 3: Life Expectancies | 43 |
| 3A: Many Kinds of Numbers..... | 43 |
| 3B Computation of Life Expectancy from Mortality Tables..... | 46 |
| 3C Analyzing World Bank Data on Life Expectancies | 50 |
| ====WRITEUP (also: attach spreadsheets in 3D) | 50 |
| 3D Histograms for the World Bank Life Expectancy Data | 51 |

| | |
|--|-----|
| 3E Variable Bin Sizes | 60 |
| 3F Probability Histograms and the CDF..... | 68 |
| Chapter 4: Reducing Data to One Number | 71 |
| 4A: Inflation: A One Number Summary of Changing Prices..... | 71 |
| 4B: The Sensitive Price Index..... | 75 |
| 4C: Composite Commodities: Laspyre's and Paasche Indices..... | 81 |
| 4D Big Data: Many Inflations..... | 89 |
| 4E Inflation as a Macroeconomic Concept..... | 93 |
| 4F Failure of Quantity Theory of Money on Australian Data | 98 |
| 5: Eugenics and the Birth of Statistics | 106 |
| 5A Malthusian Approach to Poverty | 106 |
| 5B: Sir Francis Galton: Eugenicist Founder of Statistics | 110 |
| 5C: Fisher's Failings and the Foundations of Statistics | 115 |
| 5D: Real Statistics: Alternative to Fisher's Approach | 119 |
| 5E: Contrasts between Fisher's Approach and Real Statistics: The Case of Inflation..... | 122 |
| 6: Five Quartile Summaries of Stochastic Relationships..... | 128 |
| 6A: Do the Wealthy Have Fewer Children?..... | 128 |
| 6B: Quartiles as Natural Data Summaries | 134 |
| 6C Stochastic Relationships..... | 142 |
| 6D Evolution of Distribution of Infant Mortality (IM) across time | 154 |
| 6E Comparing Progress of Countries on Infant Mortality..... | 161 |
| 6F Comparisons, Benchmarks, and Confounding Factors..... | 169 |
| 7: Probabilities, Binomials, and p-values | 175 |
| 7A: A New Definition of Probability..... | 175 |
| 7B Rules of Probability..... | 181 |
| 7C Applications of Time Branching Probability Models | 187 |
| 7D Binomial Probabilities..... | 194 |

| | |
|--|-----|
| 7E Random Sampling | 202 |
| 8: Causality and Regression Models..... | 208 |
| 8A Flawed Foundations of Econometrics..... | 208 |
| 8B Real Models Versus Econometric Models | 214 |
| 8C: Causality Defined & Compared with Regression | 226 |
| 8D: Autonomy & Invariance: Causally Defined | 233 |
| 8E: Spurious Regressions | 238 |
| 9: Assessing Association Between Two Series | 251 |
| 9A Exports & Growth Part 1 | 251 |
| 9B: Failure of Exports-Led Growth Hypothesis..... | 256 |
| 9C: The Association between Income and Consumption | 265 |
| 9D: Testing a Coin for Fairness | 274 |
| 9E: Comparing Growth Rates..... | 281 |
| 10: Some Applications..... | 287 |
| 10A Assessing the Salk Vaccine for Polio | 288 |
| 10B Discovering Causes & Differentiating them from Correlates..... | 291 |
| 10C: Do Cigarettes Cause Cancer?..... | 297 |
| 10D The GNP Illusion | 305 |
| 10E Causes of Global Financial Crisis: House of Debt..... | 309 |
| 11: Causal Analysis & Path Diagrams..... | 317 |
| 11A: Western Philosophy: Obstacle to Understanding | 317 |
| 11B: Causality as Child's Play | 318 |
| 11C: Causal Path Diagrams | 327 |
| 11D: Common Cause and Correlation | 339 |
| 11E Causation as Deep Structure..... | 350 |
| 12: Simpson's Paradox | 355 |
| 12A: An Admissions Paradox..... | 355 |
| 12B Who is the Better Batter? | 361 |

| | |
|--|-----|
| 12C Partial Mediators & Simpson's Paradox..... | 366 |
| 12D Randomization as a Solution to Confounding | 374 |
| 12E Mindless Data Crunching..... | 382 |
| PART II: Insert Title of the Part | 390 |
| Chapter Three: Insert Chapter Title | 391 |
| Chapter Four: Insert Chapter Title | 392 |
| Chapter Five: Insert Chapter Title | 394 |
| Epilogue/Conclusion..... | 395 |
| Bibliography | 396 |
| Acknowledgments..... | 397 |
| About the Author | 398 |

Chapter 0: Introduction

A: Preliminaries

Sir Ronald Fisher created an approach to statistics which has been built upon for a century to create modern statistics. This textbook advocates abandoning this approach, and starting afresh, redefining the subject matter of statistics, and developing a new methodology. This introduction provides a sketch of the ideas which led to the creation of this approach. Very briefly, rejection of Christianity as a basis for public life in Europe led to the creation of a way-of-life (*deen*) which may be termed *secular modernity*. At the heart of this *deen* is the idea that objective knowledge is possible – that we can arrive at uncontestable truths about this world, and our lives, by using reason alone. This textbook starts from the opposite premise of epistemic humility advocated by Islam: we human beings have been given very little knowledge. All the knowledge that we have is subjective and limited by the range of our experiences. We cannot arrive at universal truths using reason alone. The remaining sections of the introduction expand upon these brief remarks, but a full exposition of these ideas would require a separate book. Our current goal is only to provide some motivation and explanation for the approach to statistics taken in this textbook. We rely on the maxim that the proof of the pudding is in the eating. We believe that the approach to statistics adopted in this text will stand or fall according to how well it holds up in real-world applications, and not based on the philosophies which led to its creation.

One of the central premises of this textbook is that knowledge must be built on moral foundations. This is directly opposite of a fundamental epistemological premise of the West: morality cannot be derived from observations and logic, and hence is not part of human knowledge. The fact that moral foundations are never mentioned when conventional statistics is taught does not mean that these do not exist. The way of thinking and living known as “secular modernity” arose in the West after the rejection of Christianity. Since there is no God, afterlife, judgment, it makes perfect sense to concentrate on maximizing pleasure, power, and profits on this planet. In complex and concealed ways, these are the moral foundations on which the stock of knowledge built by the West over the past few centuries. This course was born out the attempt to replace these toxic moral foundations by those furnished by the teachings of Islam.

Where does this approach place non-Muslim students? It is an Islamic teaching that our intentions for an action determine its value. In a secular modern approach, students study statistics for personal benefits: career, wealth, knowledge. An Islamic approach is based on seeking useful knowledge for serving the creation of God, for the sake of the love of God. Any student who seeks to learn statistics for the sake of serving mankind will be comfortable with the approach adopted, since this goal is aligned with Islamic goals for pursuit of knowledge. It is worth emphasizing the secular modernity is itself a religion which replaced Christianity in the West. We do not a choice between an objective and impartial approach in contrast with an biased and subjective religion based approach. All approaches to study are based on moral foundations – we can only choose between the moral foundations of secular modernity, or those of some other religion. The ones furnished by Islam are adapted to the nature with which human beings are born, so non-Muslims would find them naturally appealing.

B: The Puzzles Posed by Human History

Looking at the broad patterns of history from an Islamic perspective raises several major puzzles. More than fourteen centuries ago, a powerful message revealed by God to mankind dramatically changed the course of world history. It transformed ignorant and backwards tribes into world leaders, and created a civilization which enlightened the world for more than a thousand years. One of the unique features of Islam was the imperative to seek knowledge: the ink of the scholars was considered as more precious than the blood of the martyrs. This led the Muslims to gather knowledge from all over the globe.

It is useful to tag 1492 as the year when the world changed radically¹. The successful voyage of Columbus to the Americas opened up vast treasures for exploitation by the Europeans. Completion of the reconquest of Al-Andalus led to the acquisition of the treasures of knowledge in the millions of books in the libraries of the defeated Muslims. This influx of knowledge ended the dark ages of Europe. The third critical event of 1492 was the purchase of the Papacy by Rodrigo Borgia, under the title of Alexander VI. This extreme corruption of the papacy eventually led to the breakup of the Catholic Church, followed by centuries of fratricidal warfare among European nations. Acquisition of knowledge, gold, and loss of morality which resulted from rejection of Christianity, led to the European scramble for wealth and power which have shaped the modern world. The amazing success of Europeans at colonizing nearly 90% of the globe, and enriching themselves while impoverishing the rest of the planet, raises several questions from an Islamic perspective:

1. What was the substance of the message which revolutionized the lives of the early Muslims, and changed the tides of history?
2. After a thousand years of success what were the causes of the decline of the Islamic Civilization, leading to the current backwardness of the Islamic lands?
3. What were the causes of the rise of West over the past few centuries, and its current global hegemony?
4. All over the globe, ignorance and poverty dominate the Islamic lands, resembling the pre-Islamic times. What is the remedy for these troubles?
5. In particular, of special importance, has the final message of God to mankind lost the power that it displayed 14 centuries ago? Can it no longer lead mankind from darkness to light?

There exists an enormous amount of literature presenting answers to these questions. A large number of movements for revival of the Islamic Civilization are based on assumptions about the right answers to these questions. The most common and widely accepted set of answers attributes the rise of the West to the development of science and technology, which placed them ahead of the Islamic Civilization on the frontiers of knowledge. Accordingly, the remedy is to acquire this Western knowledge, by developing institutions to spread education among the ignorant masses in the Islamic countries. The Ghazali Project is based on the diametrically opposite view. Western education is the source of disease in Islamic societies. It is built on toxic

¹ See “1492: The Year the World Began” by Felipe Fernández-Armesto.

moral foundations created by the rejection of Christianity. It also inculcates a worldview which glorifies the West and places zero value on the accomplishments of the Islamic Civilization. Nonetheless, a Western education also provides training in skills essential for the modern world. The only solution requires the development of an Islamic alternative to a Western education, which teaches all the skills required for the modern world within an Islamic framework. The Ghazali Project can also be thought of as a variant of project of the “Islamization of Knowledge”.

C: The Central Importance of Knowledge

The dominance of the Western civilization has spread a materialistic perspective throughout the globe. When we think of rise and fall of empires, we think in terms of wealth, power, armies, weapons, trade, etc. – that is, in terms of material causes. A far more insightful perspective emerges when we put knowledge at the center. The rise of the Islamic Civilization occurred because of a special knowledge which was given to the Muslims, as promised in the first Wahy: (God) *taught man that which he did not know*. At the center of the rise of Islam was the extraordinary emphasis given to the search for knowledge in Islam. This led the Muslims to seek knowledge from around the globe, as the lost property of the Mo'min. The rise of the Islamic Civilization can then be seen from the perspective of an increase in knowledge, as per Quran: “*My Lord, increase me in knowledge*”.

Putting knowledge, instead of the “wealth of nations”, at the center leads to an interesting and unusual perspective on history. The dark ages of Europe started with the burning of the library of Alexandria in the 5th Century, and ended with the translations of books in the library of recaptured Toledo in the 13th Century. The burning of the library of Baghdad in the 13th Century was a major blow to the Islamic Civilization. However, the real damage occurred during the colonization of the Islamic world from the 18th to the early 20th Century. 18th to the early 20th Century. Again, we should think of colonization as the conquest of knowledge. Indigenous systems for production of knowledge throughout the Islamic world were destroyed, as a matter of policy, by the colonizers. These were replaced by Western educational systems, designed to teach lessons of Western superiority and create contempt and hatred for indigenous knowledge traditions.

This is the central problem facing the Islamic Civilization. A Western education teaches us only about the knowledge created by European intellectual over the past few centuries. This automatically creates the (false) impression that this is the only type of worthwhile knowledge which exists. In particular, this implies that the knowledge given to the early Muslims, the complete and perfect revelation from God, is of no value today, at least in terms of dealing with our worldly problems. This is an illusion. However, when we make the claim that the Quran does provide guidance for today, we are bound to produce evidence for this. Currently, this evidence, an Islamic alternative to a Western education, which provides skills required for the modern world, without the accompanying indoctrination into a Eurocentric worldview, does not exist.

The challenge facing Muslims today is to devise an alternative to a Western education. This requires understanding the foundational defects in the body of knowledge created by the

West over the past few centuries. These defects stem from the rejection of Christianity, and the attempt to create knowledge without any moral foundations. Since the usefulness of knowledge can only be evaluated with reference to a goal, and useful goals depend on a conception of human welfare, a body of useful knowledge can only be built on moral foundations. Shifting from the Christian conception of welfare based on judgment and afterlife to a purely worldly conception based on pleasure, power, and profits led to a radical change in the type of knowledge which was judged to be useful. The type of knowledge created over the past few centuries by the West is extremely useful for the pursuit of wealth and power, but harmful for learning how to be the best of the creations of God. It is up to us to show that if we seek knowledge according to Islamic guidelines, this will also create, as a side-effect, beneficial worldly knowledge.

This textbook on Statistics is a proof-of-concept. It shows how the entire body of modern statistics is defective because it has been built on the wrong moral foundations. It also shows how Islamic perspectives on knowledge help us to create a superior alternative.

D: Rebuilding Knowledge on Islamic Foundations

Whoever is kept away from the Fire and admitted to the Garden will have triumphed. The present world is only an illusory pleasure." (Q 3:185)

We human beings have only been created for worship. The central question faced by a Muslim student is: how to ensure that my study is an act of worship? If we can ensure this, then the ink of our pens will have greater weight than the blood of the martyrs, on the day of Judgment. Countless verses of the Quran, many Ahadeeth, and books of scholars, testify to the central importance of pursuit of knowledge in the way of life (Deen) given by Islam. But, there is a crucial distinction between useful and useless knowledge, which has been lost from sight and vastly misunderstood, in modern times. Our prophet Mohammad SAW prayed for useful knowledge, and sought the protection of God from useless knowledge. So, when we study statistics, or any other subject, we must begin by asking what kind of knowledge this is. If it is useless knowledge, then we must stay away from it:

They learned what harmed them, not what benefited them, knowing full well that whoever gained [this knowledge] would lose any share in the Hereafter. (Q 2:102)

Knowledge which enters the heart, and brings us closer to God, is useful. Knowledge which hinders our progress towards the recognition of God, and darkens our heart, is useless or harmful. How can we learn whether or not statistics (or other subjects) are useful or useless? We must have clarity on the answer to this question prior to beginning our study.

The answer is, on the face of it, obvious. Modern statistics was created by Sir Ronald Fisher and followers, to support the vastly harmful and evil ideology of Eugenics, which held that white races were superior to others. The technique of correlation was invented to determine extent to which traits of fathers were passed onto sons. Analysis of variance was created to separate out the genetic influence from the environmental one. Can this body of knowledge, built on such foul moral foundations, be "Islamized"? The vast majority of Muslims answer "yes" – even though statistics was developed for evil purposes, useful techniques for data analysis were

developed, and these can be deployed for good purposes, as part of an Islamization of knowledge project. This textbook is built on the idea is that the answer is “no” – the techniques which have been developed are seriously deficient technically, for reasons to be explained in greater detail later. These technical flaws make it necessary for us to reject most of the developments in statistics over the past century, and rebuild the discipline from the ground up. Numerous fundamental innovations, to be described in greater detail later, are part of this new course. However, in this introductory section, we wish to address deeper issues, regarding the nature of “knowledge” itself.

Morality is the roadmap for the path towards recognition of God. Good actions advance us on this path, while bad actions move in the opposite direction. All useful knowledge is built on moral foundations. This Islamic conception is starkly opposed to the modern Western intellectual tradition, which holds that morality is not knowledge at all. Furthermore, knowledge is purely “secular” – that is, ethically neutral. From an Islamic perspective “secular knowledge” is an oxymoron: if knowledge does not teach us about the path to God, it is useless and harmful. Christianity offers exactly the same perspective about knowledge. However, historical events in Europe led European intellectuals to discard these ideas and develop a different approach. Over a century of devastation caused by religious wars between Catholics and Protestants placed European intellectuals in an impossible position. Any approach to understanding and regulating social behavior inevitably involves morality. On the other hand, sharp disagreements between Christian factions deprived them of the natural religious basis for morals. The solution that ultimately won widespread acceptance was the idea of secular knowledge -- build politics, economics, and all social sciences on purely objective and value-neutral grounds, which would be equally acceptable to all rational human beings. European intellectuals were forced to strive for an impossible goal: the creation of an ethically neutral body of knowledge about human beings and societies. They accomplished this by fraud. Moral assumptions were hidden in the framework and foundations, and labeled by misleading names such as “rationality” and “science”. This concealment created the appearance of an objective and ethically neutral body of knowledge. The name “Social Science” uses the word “science” to make a false and deceptive claim to knowledge. Similar strategies were used in the foundation of statistics, where concealed assumptions about the data were added to create the possibility of using any data to prove any preconception. Many people have taken note of the ability of modern statistics to support deception. The most popular textbook in statistics is called “How to Lie with Statistics”, which has sales more than the combined sales of all other statistics textbooks. The aphorism that “There are lies, damned lies, and Statistics” is widely known and appreciated.

The entire body of “secular” knowledge created by European intellectuals, and followers, is based on deeply flawed and concealed moral foundations. This body of knowledge must be rebuilt on the solid moral foundations provided by the teachings of Islam. This textbook of statistics is an illustration – a proof of concept. It shows how one apparently objective and neutral subject is saturated with false and misleading assumptions about the nature of human knowledge. Recognizing and removing them leads to a radically different approach to the same subject. These are grand claims, and cannot be fully supported in this brief introduction. Some indicative and sketchy arguments, together with more detailed references, will be provided in

here. We begin by noting a deep confusion created by the common use of the label “Science” for both Physical (Natural) Science and Social Science. These two domains of knowledge are dramatically different. Even though both domains require rebuilding on moral foundations, the approach to this reconstruction of knowledge will be very different for the two areas. Our concern here is not with the Physical Sciences, but with rebuilding the Western Social Sciences. However, we provide a brief paragraph of explanation for how the Physical Sciences must be rebuilt.

The modern Western Physical Sciences were created for the purpose of control and domination of the world, and they have served this purpose excellently. Islam teaches us that the worth of an action depends on the intentions. The evil intentions behind this body of knowledge are revealed in the damage and destruction they have wrought upon the planet, leading to a climate catastrophe which threatens all human life on the planet. The last century has seen the largest destruction of innocent human lives in the history of mankind, thanks to the power of the technology created and deployed for warfare. The fact that there have been some beneficial side-effects of the technologies developed for warfare should not distract us from seeing that this science has placed the power to control billions of lives in the hands of a handful of people. Rebuilding the physical sciences requires explicit recognition of the close connection between power and knowledge. Nuclear Physicists should be trained to understand their ethical responsibilities for the consequences of their discoveries. Molecular biologists who develop terminating seeds for productive varieties of food grains should understand that they are contributing to corporate profits by creating poverty and hunger. Moral foundations shape the nature of the research, and resultant technological developments – physicists who are aware of their moral responsibility would not do the research required for building more deadly bombs. The current pretense of ethical neutrality of science deceives us into separating technological developments from their use. Building all knowledge on moral foundations requires an enormously different approach to teaching the sciences. We must start from the premise that knowledge of nature is to be used for the recognition of the signs of God, and for the service of the creation of God. Models of this approach currently do not exist, and are an essential requirement for the education of the Ummah.

In this textbook, we provide a demonstration of how an alternative approach, based on moral foundations provided by Islam, lead to radical changes in both methodology and the subject matter of statistics. This is a practical demonstration – we will simply show how it is done. There is a theory which is behind the construction, but the proof of the pudding is in the eating. We will briefly discuss the theoretical foundations on which this alternative approach to statistics was constructed. First, is statistics a social science at all? Should it not be considered as a part of mathematics, which has no apparent moral foundations? To answer this question, we must consider how statistics are used in our daily lives. Upon reflection, it seems clear that statistics are used to evaluate different types of policies, and to judge relative efficacy of different actions. This evaluative role is primarily a moral one: when we decide that a certain policy is useful, we are judging between different possible social conditions according to an implicit or explicit moral framework. A number of methodological strategies are used to hide the moral framework in the conventional approach to statistics. One of these is the nominalist

methodology: according to this, we only need to look at the appearances, captured and measured by the numbers. We need not go beyond the numbers to the reality being imperfectly measured by these numbers. The “REAL” in the title of the course expresses firm opposition to this idea. We cannot understand numbers except in context of the real world; numbers cannot be meaningfully analyzed in isolation. The second toxic idea is the separation of theory and practice. The idea that the statistician can extract information from numbers and present it to the practitioners is wrong. Methods for analyzing data depend crucially on the real-world context; theory and practice cannot be separated. Whenever we use data analysis to compare and evaluate social conditions, a normative framework is inevitably involved.

We can conclude and summarize this lecture as follows. Knowledge is useful if it provides us with guidance on how to build better lives and improve human societies. This inevitable involves a normative framework for judging between better and worse. Knowledge which does not provide us with any such guidance is irrelevant and useless. Islam teaches us to seek protection from such knowledge. The entire body of “secular” knowledge created by the West over the past few centuries makes the opposite epistemological claims. It is claimed that knowledge is objective and value-free, while morality and value-judgments are not part of knowledge at all. This perverse position leads to absurd conclusions, permitting teachers to educate students the technology to build atom bombs, without accepting moral responsibility for the consequences of such knowledge. The idea that we can teach skills without discussing morality is itself a toxic moral judgment, leading to experts who deploy weapons to destroy millions of lives, without any sense of responsibility. Today, these Western educational methods are globally dominant, and result in moral stunting of students all over the world. There is an urgent need for creating an alternative. This textbook on statistics is a demonstration of how this can be done. It results from an attempt to create knowledge on an explicit moral foundation: the service of the creation of God. Having an explicit moral framework changes the methodology, approach, and the substance of Statistics, as we will demonstrate in this textbook.

E: Innovations of this textbook

In the early 20th Century, Sir Ronald Fisher initiated an approach to statistics which he characterized as follows: : “*... the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole ...*” As he clearly indicates, we want to reduce the data because our minds cannot comprehend large amounts of data. Therefore, we want to summarize the data in a few numbers which adequately represent the whole data set.

It should be obvious from the start that this is an impossible task. One cannot reduce the information contained in 1000 points of data to two or three numbers. There must be loss of information in this process. Fisher developed a distinctive methodology, which is still at the heart of conventional statistics. The central element of this methodology was an ASSUMPTION – the data is a random sample from a larger population, where the larger population is characterized by

a few key parameters. Under these assumptions, the key parameters which characterized the larger population would be sufficient to characterize the data set at hand. Under such assumptions, Fisher showed that there were “sufficient statistics” – a small set of numbers which captured all of the information available in the data. Thus, once in possession of the sufficient statistics, the data analyst could actually throw away the original data, as all relevant information from the data set had been captured in the sufficient statistics. Our goal in this section is to explain how this methodology works, why it was a brilliant contribution of Fisher at his time, and why this methodology is now obsolete, and a handicap to progress in statistics.

In the raw data, each data point is unique and informative. But Fisher's approach anonymizes all of the data by making them all equally representative of a population. This actually has parallels to our real approach – we think of the data as informing us about the real world which is hidden. But the problem is that Fisher uses an imaginary world from which the data comes, whereas we are interested in the real world. According to conventional statistical methodology, the statistician is free to make up a class of imaginary populations from which the data is treated as being a representative sample. Using this freedom, the statistician can restrict the imaginary populations to satisfy some desired prerequisite or bias. Then statistical inference will confirm this bias, making it appear as if the data is providing us with this information, when in fact, it is the bias has been built into the assumptions, and all data sets will confirm this bias.

At this introductory stage, it is hard to provide a deep and detailed discussion of all the innovations, both methodological and substantive, in this textbook. We therefore provide a bullet point list, which highlights the innovations chapter by chapter:

1. The first chapter provides a more detailed discussion of the Islamic approach to pedagogy which lies at the heart of this textbook.
2. The second chapter shows that even the simplest of operations – comparing two numbers to see which one is larger – requires considerations of the real-world context from which these numbers emerge. In contrast, conventional statistics methodology confines attention to the numbers.
3. The theme of this book is that statistics must be learnt within context of real-world applications. The third chapter discusses computation and analysis of life expectancies. It shows how assumptions go into the manufacture of numbers which are presented as objective and concrete. It also illustrates the use of some basic statistical tools like the histogram.
4. The fourth chapter discusses an issue about which conventional statistical methodology has created enormous confusion: index numbers. When objects – like universities, automobiles, research productivity – are ranked, an “index” number must be created to enable such ranking. A little-known fact is that there is no way of creating an objective index number. This means that there is no objective way of deciding which university is best, or which author has the highest research productivity, or which student has the highest overall performance. Even though rankings are done routinely, all of them necessarily incorporate subjective judgments about the relative worth of different dimensions of performance.
5. The fifth chapter provides a detailed discussion of Sir Ronald Fisher's approach to statistics, and the biases that it inherited from his racist agenda.
6. The sixth chapter illustrates how we can use statistics to compare infant mortality across time and across countries. The discussion introduces basic tools for such comparisons which differ substantially from conventional tools which are based on assuming that the data is normal.

7. The seventh chapter introduces basic probability contexts using a binomial distribution. It provides a new non-positivist definition of probability. This is different from the frequentist and the Bayesian approaches, both of which are based on positivist ideology.
8. Chapter 8 introduces causality, while Chapter 9 discusses associations (termed correlation in conventional statistics).
9. Chapter 10 discusses several real-world applications and shows how to differentiate between correlation and causation.
10. Chapter 11 and 12 provide a deeper discussion of causation, and the technical tools required for its analysis. These provide basic foundations for understanding causation, something which is not currently available in conventional textbooks of statistics. More advanced discussion of causation is left for a later book.

Chapter 1: All Human Knowledge is Built on Moral Foundations

Begin a new part here...

1A: The Illusion of Objectivity

As described in the Preface, a radical rupture between “facts” and “values” was created in the early 20th century by the emergence of the philosophy of Logical Positivism. This philosophy took the intellectual world by storm, and was used to dramatically revise the foundations of the social sciences: henceforth, they were to be fact-based and shun values of any kind. The idea that there can be value-free knowledge, heavily contested initially, became widely accepted, and forms the basis of modern Western thought. The Islamic perspective matches the pre-positivist Western perspective: all human knowledge is built on moral foundations. The central importance of the UNITY of God in Islam is reflected in the unity of knowledge: knowledge is useful if it guides us towards the recognition of God, and useless or harmful if it creates obstacles in the path towards God. This textbook on statistics attempts to illustrate this central epistemological idea by showing how it illuminates and transforms the study of statistics.

The conventional approach to statistics, adopted in all modern textbooks, assumes that statistics provide us with objective value-free knowledge about the world we live in. Indeed, it is argued that numbers are the most accurate kind of knowledge that we can have about the world. According to Lord Kelvin: “When you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind.” This dictum has two implications, both of which are false. The first is that numbers provide us with satisfactory and rich knowledge about the world. We will show that numbers can be used to conceal and deceive, and to create an illusion of knowledge about areas of ignorance. The second implication is that what cannot be measured does not exist – or at least, is not a suitable area about which knowledge can be acquired. We will show that, to the contrary, the most important types of knowledge that we can have is of those things which cannot be measured.

We must begin by understanding that the appearance of objectivity of numbers is deceptive. There is a famous book by Darrell Huff: “How to Lie With Statistics”. This is the most widely sold statistics book of all time, with more sales than all other statistical textbooks put together! In general, Statistics is strongly associated with LYING. Mark Twain said that “There are lies, damned lies, and statistics!”. Statistics appears to consist of simple facts like “ $2+2=4$ ” but it actually presents us with lies that appear to be simple truths. To understand how this magic is accomplished, we must turn to the study of rhetoric.

Statistics is a modern form of Rhetoric. Rhetoric – the art of persuasion – was at the heart of Western education for many centuries. Principles of Rhetoric go back to Aristotle. These are sometimes listed as: Ethos, pathos, logos, kairos, topos – Credibility, emotional appeal, logic, opportune moment, framing/story/narrative. However, modern rhetoric introduces powerful new

tools which were not available to the ancients. Two critical rhetorical tools are used to create the illusion that statistics is value-free knowledge:

1. The Illusion of Objectivity of Numbers – We will COUNTER this illusion by showing that: whenever we measure two or more dimensions, subjective judgments are involved
2. Separation of Theory and Practice. We will COUNTER this illusion by showing that: Whenever we go from end to end, using statistical research to solve practical real-world problems, values are involved

We now discuss how numbers create the illusion of objectivity, creating undeserved “Ethos” or credibility for statistics. One type of number is objective. For example, if I say that “There are 23 students in this classroom”. This is objective, verifiable, and the same for all observers. But if I say that the GNP in 2018 of Pakistan is 1.244 trillion dollars measured in Purchasing Power Parity (PPP), is this equally objective?

Is this second statistic of the same kind? Objective, Verifiable, Indisputable, Same for all who chose to go out to measure it? No, it is not. A large number of subjective value judgments go into producing this number. But they are concealed beneath a mask of objectivity. The World Bank data sets provide a hint of these value judgments by providing a large number of different measures of GDP, each of which reflects alternative judgments. For example, GDP can be evaluated at domestic prices, PPP prices, international prices, and many other variants. We can choose GDP numbers to reflect different types of value judgments about prosperity and welfare. In “Mismeasuring Our Lives: Why GDP Doesn’t Add Up”, Stiglitz, Sen, and Fitoussi show how current GDP numbers are based on value judgments that do not correctly reflect widely held and commonly accepted ideas about human welfare and prosperity. However, it is much easier to understand this issue (how statistics appear to be objective and indisputable, but actually conceal values) within the context of a much simpler example.

Consider the Illusion of objectivity created by College Rankings. According to the US News & World Report, they use 15 Factors to evaluate colleges. Each factor is measured by using many subfactors. To clarify the issues, we just consider three factors.

1. Graduation Rate – What % of students graduate
2. Faculty Resources – Pay/Degrees/S-F ratio
3. Selectivity – What % of applicants get admitted

Suppose a College wants to increase its ranking. What should it do? It should graduate all its students, giving degrees to those who fail or drop out. It should increase its selectivity by admitting only a small percentage of applicants. This will automatically reduce the number of students and increase the Faculty-student ratio. By taking these steps, would the quality of the college be improved? Obviously not! The quality of education is inherently unmeasurable, and all attempts to reduce it to numbers involve value judgments of many different types. This is similar to how the idea of “wealth” of a nation is subjective, and different ways of valuing different aspects would lead to very different measures of GNP.

It is worth pausing to make a very important technical point. Even though the point is simple, very few understand the full implications of this point. Whenever we do rankings – of students by grades, of countries by wealth, of colleges by quality, etc. – there is a dramatic difference between the case of one factor for ranking, versus multiple factors. For one factor – the final exam – there is a clear and unambiguous ranking. Whenever we have more than one factor, they must be combined into one number using some methodology. ALL methodologies involve value judgments that compare the two factors and assign them weights in order to combine them. A very simple example is given to illustrate this principle.

Simple Example: Consider the SAT Scores (Verbal, Math) of three students. Mahdi: (V:800, M:400) Tazeen: (V:400, M:800) Belabes: (V:600, M:600). Which of these is the BEST student? There is no objective answer to this question. Mahdi is best in English, Tazeen in Math, and Belabes has a balanced score. In different contexts, anyone of three students could be the best. The standard solution to this problem is to take a weighted average as the aggregate score: $\text{Agg} = w_1 V + w_2 M$. This leads to a unique ranking, but the choice of weights is subjective and arbitrary – this subjectivity and arbitrariness is hidden within an apparently objective single number produced as the Aggregate Score for all students.

The critical issue which needs to be highlighted here is this: every choice of weights to produce a single numerical score conceals value judgments. This concealment is damaging because it creates an illusion that there is a BEST student, when in fact the question of who is best has no answer. A second question is: which student should be CHOSEN out of three? This depends on the PURPOSE for the choice. For example, suppose we are choosing a student to compete in a debate competition. Then Mahdi would be the right choice. For a math competition, Tazeen would be the right choice. For a general knowledge quiz, Belabes would be the right choice. For a scholarship, we might want to look at which student is most in need of money, an issue unrelated to the scores. The IDEA that we can have knowledge only of things that can be measured leads to the attempt to measure things that cannot be measured (like corruption, intelligence, trust, scholarship, etc.)

The second major methodological problem with the conventional approach to statistics lies in the widely accepted Theory versus Application distinction. The narrative which statistics texts propagate is that there is a separation of tasks: The Field Expert has knowledge of complex reality and measures relevant aspects of it to produce numbers. Statistician analyzes numbers without knowledge of where they came from. In this course, we will show that this separation is not possible. Analysis of the numbers cannot be done in isolation from the real-world phenomena which led to these numbers. Statistical Analysis depends on knowing where the numbers come from, and how the analysis will be used in real-world context. For example, the choice of weights we use to aggregate SAT Verbal and Math scores into a single number for ranking students requires us to know the purpose for which the rankings are being done. The adjective “real” in the title of the course indicates a methodological approach where numbers are always analyzed together with the real-world concepts which these numbers are supposed to measure. We reject the conventional approach to statistics which may be labeled as “nominal”

statistics in opposition to “real”. The conventional approach is based on the analysis of numbers by themselves, without any reference to their real-world origins.

The central pillar on which Western social science is constructed is a claim of objectivity: knowledge is FACTUAL and VALUE-FREE. The goal of this course is to show that this claim is wrong, at least in the context of statistics. Values are inevitably involved in all aspects of relevant and useful human knowledge. Since the generation of measurements and data involves effort, it must be done for some human purpose. All human goals reflect an orientation towards life and embody values. In statistics:

1. Values lie in the construction and selection of numbers: measures tell us that what is being measured is worth measuring.
2. Values lie in the field of application, and the purpose for which statistics are being gathered and analyzed.

When we go deeper into the numbers, to explore where they come from, and what real-world problems we can solve by using them, we will show that the moral framework of Islam provides essential guidance, which differs substantially from the values implicitly embodied in the current Western approach to statistics.

1B: Purpose: The Heart of An Islamic Approach

In the previous lecture, we explained that statistics is NOT Objective and Factual, even though it claims and pretends to be so. What we consider worth measuring, and how we go about measuring it, reflects value judgments:

1. Most numbers that we use are “INDEX numbers” which combine multiple factors to get one overall measure of a complex multidimensional phenomenon. Whenever we combine factors, subjective value judgments are involved in the selection of factors, and in assigning relative importance (weights) to these factors, in order to come up with a single numeric measure.
2. Whenever we APPLY statistics to the solution of real-world problems, real-world context calls for value judgments. This value-laden aspect of statistics is hidden because the standard approach separates theory from practice. In these lectures, we will show that this separation cannot be done, because the theory we use depends on the real-world application. When we do theory and practice together, Islam provides us with guidance on the values required.

At the core of an Islamic epistemology is the idea of the unity of knowledge, which is in sharp contrast with the duality at the core of Western epistemology. In contrast to the idea of a sharp separation between facts and values, facts acquire importance (and become worth

measuring) only if they are relevant to the achievement of our life-goals. From this perspective, the study of statistics must be framed within the context of the following questions:

1. Why are we studying statistics?
2. How will this knowledge help us to achieve our life goals?
3. What are our life goals?

These questions arise naturally for all students asked to devote significant amounts of time, effort, and money to the study of statistics (or of any branch of knowledge). Why does the western intellectual tradition fail to discuss these questions? Reasons are deep, coming from the assumption of the meaninglessness of life itself. The following poetic quote from Bertrand Russell explains a fundamental background assumption that human life is a product of chance, and without any purpose or meaning:

MAN ... his origin, his growth, his hopes and fears, his loves and his beliefs, are but the outcome of accidental collocations of atoms; that no fire, no heroism, no intensity of thought and feeling, can preserve an individual life beyond the grave; that all the labors of the ages, all the devotion, all the inspiration, all the noonday brightness of human genius, are destined to extinction in the vast death of the solar system, ... Only within the scaffolding of these truths, only on the firm foundation of unyielding despair, can the soul's habitation henceforth be safely built.

The Quran (92:4) tells us that “Surely the ends you strive for are diverse”. Different people have different goals, and therefore acquire different kinds of knowledge, relevant to their goals. There is widespread consensus on the idea that meaningless lives are not worth living. In this course, we will adopt an Islamic perspective on the purpose of life, which is the opposite of that of secular modernity. Those who consider life to be meaningless are not part of the intended audience for this course: if life itself is meaningless then all human effort and knowledge is also meaningless. However, non-Muslims who do find life meaningful, and strive for higher goals than the pursuit of pleasure, will find substantial common grounds with an Islamic approach to knowledge.

Islam firmly rejects the idea that the Creation is without purpose, and that human life is without meaning. Islamic teachings clearly specify the purpose of our lives:

Say: Lo! my worship and, my sacrifice and my living and my dying are for Allah, Lord of the Worlds. (Qur'ān, 6:162)

Islam teaches us that human lives are very precious: [Quran 5:32] “whoever saves a life is as though he had saved all mankind”. Human beings are the best of the creations, and have a huge amount of potential, in terms of capabilities that can be developed. History furnishes with examples of human beings who have achieved spectacular excellence, and also those who have been spectacularly evil. How we answer the question of “how we can develop our potential to the maximum?” determines how we spend the brief and precious moments that have been given to us. Any use we make prevents us from using these moments in other ways. Thus all human

choices – to acquire knowledge (about statistics) or to choose not to do so – are dependent on the goals of our lives, and whether or not we can use this knowledge to achieve our goals.

The Islamic Approach to knowledge distinguishes strongly between Useful and Useless Knowledge. Our prophet Mohammad SAW taught us to pray for and seek useful knowledge, and to seek the protection of Allah from useless knowledge. This distinction makes sense when there is a purpose to life. Knowledge which is helpful to achieve our life goals is useful, while that which does not help is useless. So, the natural question is, “how can we study statistics to make it useful knowledge?”. If we can do so, then the ink we use will be valued like the blood of the martyrs, the efforts we make in learning will count like the greatest Jihad.

For this purpose, we must make the highest intentions: the Service of Mankind for the Love of Allah. How can learn statistics to serve mankind, for the love of Allah? Note that Statistics is used routinely for deception. Examples are given in “How to Lie with Statistics” & “Confessions of an Economic Hitman”, which show how data can be used to deceive. So we must learn statistics to learn to distinguish truth from lies, and to protect humans from harms caused by false and deceptive statistics. In contrast to deceiving ourselves and others by numbers, Islam offers us a distinctive alternative. Allah T'aala Himself asks us to seek knowledge: “Rabb-e-Zidni Ilma”. Many Hadeeth teach us about the value of knowledge. Those who seek knowledge are given higher status than warriors and pious people. The thirst for knowledge was the driver for the amazing rise of ignorant and backward Bedouin to world leaders, and the creation of a dazzling Civilization that lasted more than a thousand years. The loss of this thirst for knowledge is responsible for our current lowly condition.

The powerful knowledge, which can transform our inner lives and the external world, requires struggle. Many Ahadeeth on the rewards for seeking knowledge are there to produce high motivation for the Great Effort Required, Allah T'aala says in the Quran that ‘Those who struggle in our pathways, we will guide in our pathways’. This shows us an action-oriented approach to learning. It is in the process of struggling to use knowledge to create positive change, in our lives, and in the world, that Allah T'aala will give us the understanding we need. In particular, neutral detached observation is not permissible. Allah T'aala asks us to struggle with our hands, and with our tongues, and with our hands, to remove evil and to create good. It is important to note that we have full responsibility for the knowledge that we acquire: whether it is used for good or evil. So unlike that traditional theory/practice divide, the Muslim statistician must be concerned with the issue of whether or not his data analysis will be used for good or bad purposes.

The greatest motivation for the effort to acquire knowledge comes from the understanding that We (human beings) have been created with the greatest potential: to be Ahsan-e-Taqweem (best of the creations), and rise higher than the angels. Every human life is (potentially) worth the entire humanity! How can we realize our hidden capabilities to the maximum? This involves acquiring the knowledge and life-experiences which will bring out these capabilities. Some steps in this direction are discussed in related posts linked below:

1. Recognizing our true identities: [Learn Who You Are:](http://bit.do/azwy) bit.do/azwy

2. [Reaching Beyond the Stars](#): Aim high to get great results. bit.do/azrbs

1C: Knowledge: Islam vs. West

Even though we are unaware of it, the ideas of philosophers have dramatic effects on lives of common people. The little fish does not know why there are billions of tons of plastic in the ocean, but its life is dramatically affected by this plastic. The story of how the conception of knowledge changed in the 20th Century is deep and complex. We will drastically over-simplify to provide a quick introduction to the basic ideas we need to know, for our present purposes.

The idea of “knowledge” was RE-DEFINED in 1930’s by Logical Positivism (LP). Modern Social Sciences were built in early 20th Century according to these new conceptions of knowledge. LP was rejected by philosophers around 1960’s. (For more details, see [The Emergence of Logical Positivism](#)). However, the news that LP has been rejected did not travel to social scientists. In particular, our fields of study (Economics, Econometrics, Statistics) are still based on Logical Positivist foundations. These foundations are wrong, and need to be replaced. The “Islamic approach” taken in this course will rebuild the discipline of Statistics on new, non-positivist, foundations.

We are ALL trained to be logical positivists, without our knowledge. We learn our “theory of knowledge” not from discussions of philosophy, but from what we are taught. The syllabus, the subjects we are taught define for us the scope of knowledge. This syllabus has changed radically over the 20th Century. As [Julie Reuben writes in “The Making of the Modern University: Intellectual Transformation and the Marginalization of Morality”](#), college curriculum today has been shaped by the influence of logical positivism. Early 20th Century colleges sought to build character, but later this quest was abandoned, because “morality” is unscientific and meaningless according to positivism.

We must briefly discuss how the Logical Positivists understand knowledge. Very briefly, according to LP:

1. Scientific Knowledge is based on objective observations of external reality – these are facts.
2. Knowledge of our internal subjective experience is not part of “scientific” knowledge.

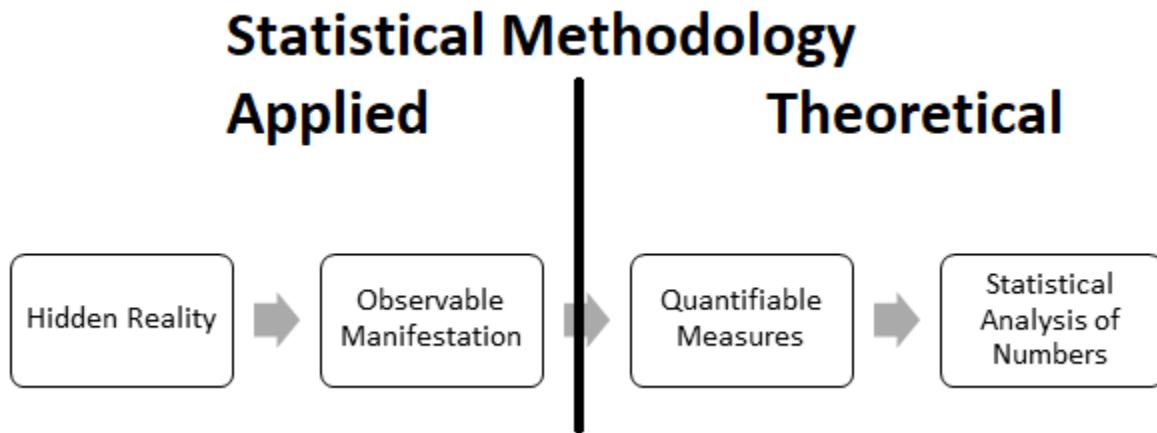
(For more details, see [The Search for Knowledge](#)). This is a MIS-UNDERSTANDING of science – Science is also based on guesswork about the hidden mechanisms, like gravity, subatomic particles and so on. But this misunderstanding became dominant and widespread in European intellectual tradition because of loss of faith & rejection of Christianity. This created a distrust of unobservables which is reflected in the popularity of logical positivism.

Contrary to what we are trained to believe, modern Social Sciences (mSS) were born in the early 20th Century. They do not go back to Adam Smith or earlier sources. The foundations of mSS are based on exclusion of heart, soul, morality, from knowledge. In particular, Life experience is the most important source of knowledge, but Life experience is NOT scientific, and hence not knowledge, according to LP.

Modern Social Sciences were creating by rebuilding knowledge by excluding the heart. For example, in economics, the hidden preference of the heart was replaced by the observable choice that it leads to. In psychology, Skinner replaced hidden internal states of the psyche by external observed behaviors (Behavioral Psychology). This led to the conception of man as robot, programmable by stimulus response. Accordingly, his book has the title: Beyond Freedom and Dignity.

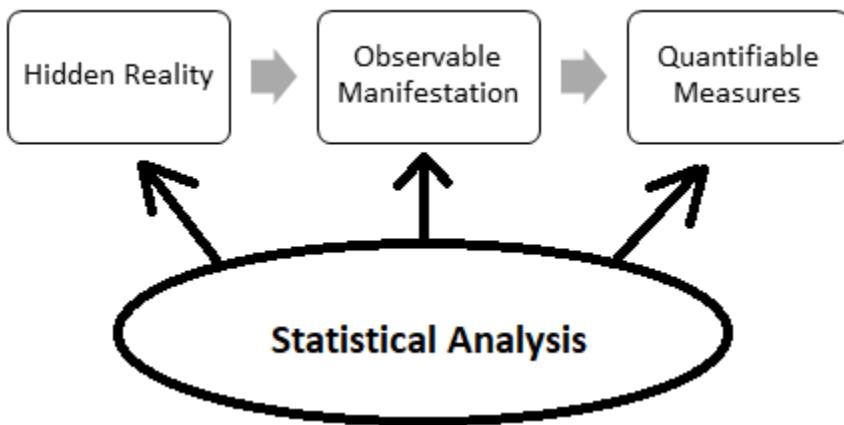
For the Islamic perspective, the Quran talks about the TWO journeys: Believers go from Darkness to Light, while non-believers go from light to darkness (2:258). To understand this verse, we must differentiate between internal and external knowledge. Whereas the West has made tremendous advances in knowledge of external reality, it has also lost understanding of the inner nature of man. Just as the progress is manifested by amazing advances in weapons, technology, medicine, and scientific knowledge, the loss of knowledge is manifested in drugs, crime, suicides, perpetual warfare, and breakdown of society cause by the breakdown of the fundamental unit of the family. (See [Origins of Western Social Science](#))

Coming back to our topic, how does this emphasis on observables manifest itself in statistics? Logical Positivism is at the heart of current methodology of statistics which creates a sharp separation between theory and practice, as demonstrated in the following diagram.



The APPLIED field expert deal with hidden realities and observable manifestations. Statisticians deals with measurements and quantifications – numbers only. The idea is the science can only deal with observables, and statisticians can only deal with numbers which measure the observables. These ideas create major difficulties in real life when dealing with qualitative phenomena like intelligence, quality of research, dynamism, creativity, persistence, etc. which cannot be reduced to numbers.

In contrast, an Islamic approach requires an INTEGRATED analysis of the statistics together with the hidden real world phenomena which generate the numbers that we see:



We cannot separate the theory from the practice. An Islamic approach to statistical analysis requires consideration of all three – the hidden reality, the observable manifestations, and the quantifiable aspects of the observables. (See also, [My Journey from Theory to Reality](#))

Even though positivism has been rejected by the philosophers, it continues to rule the hearts and minds of social scientists all over the world. Unfortunately, this has led to a lot of harm. Some simple examples will be mentioned for the sake of illustration.

1. Management by numbers: “You cannot manage what you cannot measure!” These widely accepted words of wisdom are deadly. When we start counting the number of articles published, faculty starts publishing to increase count, instead of producing high quality research.
2. The US defeat in Vietnam War was caused by applying heartless quantitative methods to increase casualties among the Vietnamese without considering the human effect of increasing hatred and unity against the invaders.
3. Many studies show that for most kinds of work, measuring output and monetary rewards do not work well, because human motivation is driven by other factors.

So, what does this mean in terms of prescriptions for Islamic approach to statistics? It is essential to realize that Statistical Theory cannot be studied in isolation from real world. Abstract theoretical concepts can only be understood when their application is demonstrated and differentiated in real world context. As a specific example, consider the Mean, Median, and Mode, which are commonly used measures of central tendency. Students study the theory without learning What it means to use one instead of the other. To decide which one is better requires knowing the PURPOSE of the analysis, the nature of the data set, and the relation between measurement, observations, and the real world. It is only in context of a real world application that we can understand these theoretical concepts. This is what we hope to demonstrate in this course.

Final Words: It is narrated in a Hadeeth that useful Knowledge enters the heart. When knowledge is related to real world, then it becomes connected to human lives. Then knowledge can be harmful, irrelevant, or useful, according to its good or bad impact on human lives. When it

relates to lives of other human beings, it also relates to our lives. By relating knowledge to life experience, we make it come alive for students. This requires overcoming the theory/application barrier common in standard approach to statistics.

1D: Islamic Pedagogy: Engaging the Heart

Islamic education requires simultaneous work on Heart and Head; both are essential to the learning process. This contrasts with the Western approach, which is purely “rational”, engaging the head but not the heart. The two dimensions of effort required can be briefly summarized as follows:

1. **Mental Efforts:** Struggle with material to understand it. Explain ideas to fellow students, or others. Absorb & articulate, in your own words
2. **Spiritual Efforts:** Make Dua's for the opening of our hearts to knowledge. Understand the importance of knowledge. Make the intention to use knowledge of service of mankind, for the love of Allah.

Now we discuss these two dimensions of the effort in greater detail.

Spiritual Efforts

Tazkiya (Purification of the Heart): My worship, struggle, living, dying, all are for the sake of Allah alone – In particular, I am acquiring knowledge to seek the pleasure of Allah (by fulfilling his command to seek knowledge) and out of the love of Allah.

An Attitude of Humility: We have no knowledge except that which YOU have given us. As the angels said to God (La Ilma Lana, illa ma Allamtana). Similarly, Ayatul Kursee teaches us that We can have no knowledge except fragments which Allah T'aala chooses to grant us. Wala Yuhitoon b' sheyim min ilmahi illa be mashaa'.

Thus we must seek useful knowledge from Allah with sincerity and humility. In the process of doing so, it is ESSENTIAL to avoid WRONG intentions for knowledge:

1. Do not seek knowledge to demonstrate your superiority over others!
2. Do not seek knowledge to ARGUE with the ignorant.
3. Do not seek knowledge for popularity, to entertain audiences.

The hellfire is for those who do this. Fame, Pride, Popularity – all are intentions to use knowledge for pleasure of our own EGO. Seeking to SERVE human beings so as to become popular and praised is NOT acceptable. Remember the famous Hadeeth about the Alim, Shaheed, and Generous Man, who will be thrown into hellfire because of lack of sincerity – they sought popularity, fame, and got it, but there is no reward for them in the Akhira.

Sincerity means serving the creation of God *for the sake of the LOVE of Allah*, and NOT for popularity, for earning their gratitude, or for receiving their recognition, respect, or rewards for our service.

Another aspect of spiritual training is the filling of our hearts with DESIRE for knowledge. To do this, contemplate that Allah T'aala introduces Himself as the One who gives knowledge. The Prophets were all teachers of mankind, taking them out of the darkness towards the Noor of Allah. Knowledge is among the greatest of the treasures of Allah. Angels ordered to prostrate before Adam AS AFTER he was given knowledge of the names. That is knowledge is what makes mankind the Best of the Creations. In order to create the desire for learning, the Prophet SAW explained that reciting one Ayah is more desirable than getting free camels. Remember that camels were “wealth” and the highest status symbol among the Arabs. The Prophet encouraged the companions to prize knowledge about all worldly possessions. This is also what we need to learn.

The Mental Effort

It is the desire for learning which creates the energy required for the STRUGGLE to understand, necessary to acquire knowledge. Some aspects of this struggle are discussed below.

To learn knowledge in books, or that given by a teacher, we must absorb this in our own minds, and be able to re-create it. This process transfers the “ownership” of the treasure of knowledge contained in books and teachers to our minds. Allah T'aala instructs the Prophet to take the time required in reciting (and understanding) the Quran:

20:114 High above all is Allah, the King, the Truth! Be not in haste with the Qur'an before its revelation to thee is completed, but say, “O my Lord! advance me in knowledge.”

To acquire understanding, we must contemplate, and then re-express ideas in our own minds and hearts.

Another essential step in the struggle to understand is “Acting upon knowledge”. We must seek opportunities to use knowledge. The simplest opportunity is provided by the exercises which require us to put into practice skills that are being taught in textbooks. But more generally we should seek opportunities to put our knowledge into practical use in any aspect of our lives where it seems relevant. This also puts the burden on the teacher to ensure that he/she teaches USEFUL knowledge, which has relevance to our life experiences.

In the process of translating knowledge into action, it is essential to understand the difference between doing and understanding – experiential and intellectual knowledge. Doing and Understanding are SEPARATE kinds of knowledge, and we need to learn both kinds. To see the difference, consider a simple example of adding fractions. Suppose we want to learn how to add $1/2$ and $1/3$.

DOING: To add $1/2$ and $1/3$, we can apply the rules. First Multiply denominators $2 \times 3 = 6$. Then cross multiply numerator and denominator to get $3/6 + 2/6 = 5/6$. However, this rule gives us NO understanding of WHY this works, or WHAT it means to add the fractions. That requires a SEPARATE effort.

UNDERSTANDING: To understand the concept of fraction addition, it is useful to explain it in a context that is ALREADY known and familiar to the student. Generally speaking,

we expect students to have experience with round cakes, pizza, or circular pies. With this familiar context, we can create understanding by asking students to divide the pizza into six equal slices. Then it is clear that One Half is 3 slices, and On Third is 2 slices. So the sum is clearly 5 slices, leading to $1/2 + 1/3 = 5/6$. This is understood because it is explained within the context of life experiences familiar to the student. Note that even if the student understands this perfectly, he/she will NOT understand from this the RULE for adding fractions, and maybe confused if given different fractions to add. Experience accumulates slowly, and after a lot of experience, you can arrive at the rule which covers many different special examples into a big picture.

Another essential aspect of the struggle to understand is to Articulate and Explain the knowledge you have to others. We must be eager to share our treasures, to spread knowledge to others. This requires motivation, articulation, and explanation. Remember that Muslims are lifelong learners; we are taught to “Seek knowledge from cradle to grave”. The Prophet SAW was eager to spread his message to all of humanity – so much so that Allah T'aala consoles him in the Quran, telling him not to kill himself with sorrow if they do not listen to you. On our part, we must learn to VALUE other seekers of knowledge. Give them respect, attention, recognition. Give them time, be patient with them, and encourage them in the difficult process of learning. The Quran teaches us to make room for newcomers in our circles of education.

To summarize, Islam is unique among religions in giving a prominent place to knowledge. The angels prostrated to Adam AS *after* he was given knowledge of the names by Allah T'aala. Ibraheem AS chides his people for following inherited customs and beliefs without thinking about them. This unique status of knowledge is accompanied by a unique epistemology, as well as a unique approach to teaching and learning. These approaches, which have no match to the Western intellectual tradition, are described more explicitly in the next section.

1E An Islamic Methodology and Epistemology

When we consider within its context as a branch of human knowledge, it becomes immediately apparent that this is a branch of modern rhetoric. Looking at Statistics as a method of persuading with numbers gives much greater clarity in terms of how we need to study statistics. Any attempt to persuade is based on some value-based positions. Understanding this connection is required to answer the objections to our approach created by the secular modern perspective. Conventional statistics assumes that numbers are purely objective measures, and can be analyzed without reference to value-laden objectives, and the narratives around the numbers required for rhetorical persuasion. Some frequently raised objections arising from this perspective are listed and answered below.

1. It is entirely possible to present all the statistical and technical material within a secular, materialistic, framework, aligned with atheism and modernism.
2. It is also possible to re-create this approach within Christian, Hindu, Buddhist, Confucian, or any other philosophical or cultural tradition.

To those trained in the secular modern intellectual tradition – and this is everyone with a Western education – the above observations lead to deeper puzzles.

3. What is the need of using an Islamic methodology and epistemology? This terminology restricts the audience, repels many potential students, and, seemingly, offers no obvious advantages. Can we not avoid inserting values and morals into a neutral subject like the study of numbers?
4. FURTHERMORE – it appears dishonest – packaging of modern mathematical subjects within ancient frameworks unrelated to the topics under discussion.

A DECEPTION created by Modern Secular thought is that we have an OPTION to keep values out of knowledge. In fact, Modern Secular Thinking is ITSELF a RELIGION, which presents itself as a neutral, unbiased, and objective. There are HIDDEN NORMATIVE assumptions upon which the entire structure of Western Social Science is built. These assumptions TELL us about the purpose of life, without explicit articulation.

In the realm of statistics, most numbers are claimed to be objective, but are full of subjective assumptions and moral values built into foundations. This appearance of objectivity, and claim of neutrality, creates a powerful deception. This enables HOW TO LIE WITH STATISTICS, and Economic Hit-Men. Widely believed deceptions which drive policy all over the globe are: College Rankings, Student Evaluations via Objective Quizzes. The most important deception is the use of the GNP per capita as a measure of development. The use of this number represents a SUBJECTIVE opinion about how to measure prosperity and wealth, and the hidden values embodied in this number drive economic policy all over the planet.

Before approaching any subject matter, The First Question we must ask is: WHY? What is the PURPOSE of studying statistics? In the modern secular approach, this QUESTION is BYPASSED. BUT this question CANNOT be bypassed. PURPOSE is IMPLICIT, HIDDEN in the way the subject is presented. When we ask about the PURPOSE of Study, we must ask about Purpose of LIFE itself. ESSENTIAL QUESTIONS are:

What is the purpose of my existence? How will study of statistics HELP me in achieving my life goals? Our “objective” approach hides the important fact that All Knowledge COMES from LIFE EXPERIENCES – these are distilled into a form which make them appear objective. Useful Knowledge MUST relate to LIFE EXPERIENCES. But no mention is made of LIFE and of EXPERIENCE in conventional approach – WHY? These questions were at the heart of the Wisdom of the Ancients, which has been Forgotten by the Moderns:

1. Socrates: “To know thyself is the beginning of wisdom”.
2. Socrates: “A Life Un-examined is NOT worth Living”
3. Aristotle: KNOWLEDGE come from knowing WHY?

Learning why involves learning about FOUR Causes, which are best illustrated by an example. Consider a TABLE in the dining room:

1. Material Cause: Wood – material used to create table
2. Formal Cause: The FORM or DESIGN with four legs, flat surface, raised.
3. Efficient Cause: Carpenter who took materials & shaped it into design.
4. FINAL CAUSE: the PURPOSE of creating a table.

It seems OBVIOUS that Knowledge MUST be related to PURPOSE. But conventional approach DOES NOT discuss Purpose: WHY? This is because of the HIDDEN ASSUMPTION of Modern Secular Thought: There is no purpose to life. One of the leading atheists and influential philosophers of the 20th Century, Bertrand Russell, expresses this explicitly in his essay on “A Free Man’s Worship”:

That man is the product of causes which had no prevision of the end they were achieving; that his origin, his growth, his hopes and fears, his loves and his beliefs, are but the outcome of accidental collocations of atoms; that no fire, no heroism, no intensity of thought and feeling, can preserve an individual life beyond the grave; that all the labors of the ages, all the devotion, all the inspiration, all the noonday brightness of human genius, are destined to extinction in the vast death of the solar system,

According to the modern secular approach: All human effort is MEANINGLESS – No purpose to life, and no purpose for knowledge. Islamic Teachings Strongly Reject This

5. (75:36) Does Man think he was created without purpose?
6. (67:2) (HE) created death and life (as a trial) to see who is the BEST in deeds –
7. (51:56) We have been created only for WORSHIP.

Islamic Approach teaches us HOW to study statistics as an act of worship. How to make our search for knowledge, the best of deeds, so that the ink of the scholars becomes as precious as the blood of the martyrs? This is REQUIRED for us as Muslims. We have NO OPTION but to CREATE an Islamic Approach. This would have happened naturally, but Islam came as a stranger, and has become a stranger. The process of colonization has nearly destroyed the Islamic Intellectual traditions and heritage. As a result, a GREAT EFFORT is required to create a new approach to modern subjects, aligned with the traditions of Islam. We have no option but to reject the Secular Modern Approach: Life is NOT Meaningless. We CAN create alternatives based on OTHER meanings. All religions and philosophies prescribe MEANING, and make search for the Meaning of Life the MOST IMPORTANT problem. All other problems are SUBORDINATE to this. The search for meaning in our lives is CLOSELY related to IDENTITY: Who Am I? How can I approach the study of statistics so that it enhances my life experiences, and helps me learn who I am? According to the Islamic Approach: We acquire knowledge to RECOGNIZE the SIGNS of GOD. HE is hidden in the wonders of the world around us. We can also recognize HIM in our internal world, studying our own selves leads to a recognition of God. Islam provides us with a high and inspiring purpose for our lives. The BEST of DEEDS: To serve mankind PURELY for the sake of the LOVE of Allah, without intention for fame, popularity, recognition, reward, thanks.

While we must reject the modern secular approach because it says that life is meaningless, there can be Alternative Approaches to meaning, different from the Islamic approach. We must SPECIFY PURPOSE (for our lives, and for the role knowledge must play in

our life experiences). Then we must Adapt our study of statistics to that purpose. Alternatively, we can SEARCH for PURPOSE. The Modern Secular Approach CLOSES the opportunity for meaningful discussion on the most important questions we all face:

1. HOW can I lead a meaningful life?
2. HOW can I learn WHO I AM?
3. HOW can I know of the HIDDEN potential buried within me?
4. HOW can I develop my talents for EXCELLENCE?
5. In WHICH direction should I develop my capabilities?
6. WHAT should be my priorities in LIFE?

It is a CRIME to teach students subjects WITHOUT discussing these Central Questions. Failing to discuss these questions leads to wasted lives spent on futile efforts on irrelevant goals.

Concluding Remarks: Purpose is MISSING from standard approach. Islam provides a clear-cut purpose which is DIRECTLY opposed to the HIDDEN message of the secular modern approach: meaninglessness of life and of knowledge. Conventional approach makes no distinction between useful and useless knowledge, because all effort is meaningless. When we study statistics with a PURPOSE, this actually CHANGES the subject: we cannot study numbers in abstract, theories without context. Statistics is based on creating narratives supported by numbers. The essential role of the NARRATIVE is NOW coming to be recognized. For convincing narratives, good rhetoric, numbers must be studied WITHIN the real world context from which they originate. This STRONGLY differentiates a PURPOSEFUL Islamic approach from the modern secular approach.

Links to Related Materials

1. [Learn Who You Are](https://bit.do/azwya): The search for identity. bit.do/azwya
2. [Reaching Beyond the Stars](https://bit.do/azrbs): Aim high to get great results. bit.do/azrbs
3. [How to Inspire and Motivate Students](https://bit.do/azhims). When we relate our subjects to life experiences, and to a great vision for serving mankind, it becomes possible to inspire and motivate students: bit.do/azhims
4. [The Ways of the Eagles](https://bit.do/azwoe) : Education teaches us to think low, like crows. Instead, learn vision & purpose to soar the skies, like eagles: bit.do/azwoe
5. [Three Mega Events Which Shape Our Thoughts](https://bit.do/azgt4): How to free our minds from chains created by powerful historical forces: bit.do/azgt4

Chapter 2: Comparing Numbers

The first chapter introduces the topic to the reader. As an example, the title of the example book's first chapter is "Understanding Your Child." Start off providing a brief overview of what the chapter contains and then transition smoothly into your supporting points. Try to

keep the language simple and understandable to generate rapport with the reader and keep them engaged.

One way to generate rapport with readers is to start your chapter with a quote from a famous person. Make sure the quote is relevant to the chapter topic so that you can use it later to illustrate the key points made in the chapter. If you begin the first chapter with a quote, stay consistent and begin every successive chapter with a quote.

Alternatively, you can begin a chapter with “Did you know ...?” Follow up with some statistics that most people may not be aware of. You can apply this technique in different sections of your book and not just in your introduction. Asking the reader questions (sometimes rhetorical) is a good way to keep them engaged and often entertained.

Another opening strategy is to ask the reader to imagine a specific situation. For example, “Imagine a world where every single child is trained to ...” The goal here is to draw the reader in from the get-go and hold their attention.

The key to starting each chapter is to do so consistently throughout.

2A Comparisons Driven by Purpose

The simplest statistical operation is to compare two numbers and decide which is larger. If we look at the numbers only, then this is completely TRIVIAL. For example, here is World Bank data on GDP for Pakistan and Morocco in 2009:

| Country | GDP in Billions of Dollars 2019 | GDP/capita in Dollars 2019 |
|----------|---------------------------------|----------------------------|
| Morocco | 119 | 3345 |
| Pakistan | 284 | 1388 |

Which country has higher GDP? Gross GDP of Pakistan is much higher than that of Morocco, but GDP per capita of Morocco is much higher than that of Pakistan. So what can we conclude from this comparison? Nothing, unless we go beyond the numbers to ask WHY we want to make this comparison.

Four Questions will guide our study of statistics. These questions are not part of conventional approach to statistics.

1. Why are we comparing numbers?
2. What do the numbers MEAN?
3. How were these numbers computed?
4. What will be done with the analysis?

Instead of thinking of statistics as the analysis of numbers in isolation from the real world, we will consider statistics as a branch of *rhetoric*: how do we use numbers to make arguments, and to persuade?

When asked to compare GDP to decide “Which country is “Wealthier”?”, we need to go into the background. WHY are we asking this question? Depending on the why, different analyses will be appropriate. In case of WAR between the two countries, gross GDP matters – this indicates the economic resources available to the country – this was the original context of “Wealth of the Nations” by Adam Smith. If we want to compare living standards, GDP/capita matters. This shows how Statistical Analysis CANNOT be separated from real world context and purpose of analysis.

When we want to think about the arguments being made with numbers, we have to look at the historical context. Even though Adam Smith introduced the concept of the “Wealth of Nations” in 1776, measure of GDP were first introduced by Simon Kuznet’s in 1934. Why this long delay? To understand this, it is important to realize that History is the conquest song of the victors. The measure of development is defined by the powerful to make them look good. There is no entry for “democracy” in 1930 Encyclopedia Britannica; the British were an aristocratic society. For more on how the concept of development, and corresponding measures, have changed over time, see: “What is Development?” – bit.do/azwid

When we think about the meanings of the GDP measure, several possibilities emerge:

1. Measure of POWER?
2. Measure PROSPERITY or MATERIAL WELFARE?
3. Measure of HAPPINESS?
4. Measure of fulfillment of basic needs?

We have considered the possibility of measuring national power by GDP. But if this is the purpose, we should also consider military power, political power, and other factors which contribute to national power, in addition to the economic power which may be measured by GDP. Similarly, for the other three possible meanings, in each case, GDP is seriously deficient.

The question of “How can we compute BETTER numbers?” is considered in depth by the Stiglitz-Sen-Fitoussi Report on GDP. We briefly summarize some of the key issues presented in this report.

They explain that we need to study deficiencies of the GDP measure because GDP/Capita is the CENTRAL measure of National Wealth. This measure DRIVES economic policy across the globe. Flaws in this measure lead to flawed policies. They establish that there ARE very serious flaws in GDP. One way to see this is to note that measures of inflation, growth, prosperity from national income accounts, differ drastically from public perception of the same phenomena. There are many reasons for this. One of the important reasons is that the statistical AVERAGE does not reflect the experience of the masses. For example, high inflation in food prices will hurt the poor masses, and make them experience a very high inflation. However average inflation may be low if non-food commodities are not changing in price.

Another question which statistician do not ask is: “What will we do with this analysis?”. Stiglitz-Sen-Fitoussi consider the policy implications of changing the GDP measures. They say that the Global Financial Crisis occurred because we were NOT measuring the right numbers. Similarly, the Environmental Crisis, Looming Destruction of Human and Animal Habitat, is happening because the GDP numbers do not reflect these concerns. Better measures of human prosperity would lead to better policies. So it is vital to go beyond the narrow focus on numbers, and look at the four questions posed earlier.

The Stiglitz-Sen-Fitoussi report ends up with the following **Recommendations**

5. Shift from Production to Well-Being
6. Look at Income & Consumption, rather than production.
7. Emphasize Household Perspective, breakdown by income categories.
8. Broaden Wealth to non-market activities.
9. Improved measures of health, education, environment.
10. Improved measures of social networks, family, friends, community

All of these changes would radically affect the standard GDP measure. When we analyze GDP without taking into account what it means and how our analysis will affect the future, we become part of the chain of ignorance and neglect which leads to major disasters on economic and environmental fronts.

Concluding Remarks: We have seen that connecting Numbers to Real World Applications requires going far beyond standard scope of statistics. From the Islamic point of view, every small participant in an act gets credit or blame for the whole act. So, the statistician CANNOT escape responsibility for how statistical analysis is used. THEREFORE, we MUST broaden the scope of study to include questions discussed earlier, beginning with the WHY of analysis. We cannot say that I will just analyze numbers, let the others USE my analysis for whatever purpose they want to use it. This leads to End-To-End analysis – we start with a real world question, consider how to answer it, and use statistics for solution if it is required. Statistical analysis is a means to an end, and not the goal in itself.

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are always related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 can be “How Well Do You Know Your Child?” You can use a real-life story and talk about the importance of knowing your child.

2B Arbitrariness of Rankings

One of the best sources of learning is reading articles and books. Good articles and books encapsulate deep wisdom, which authors have gathered from their life experiences. Ultimately, the only source of knowledge is life experience itself. Since we have only one life to live, we can only gather a small amount for ourselves. Reading gives us access to the fruits of the life experiences of millions of scholars, throughout the centuries of written works. It is essential to be selective in this reading. This is because the amount of false and misleading information is vastly greater than that which is useful and relevant. Furthermore, even the useful and relevant material is so extensive that we will only have a chance to read a very small portion of this in our entire lifetime. One of the tasks of a good teacher is to provide guidance in this regard. Having read thousands of articles, select the few that stand out for students. If the teacher can point the student to one article that summarizes the wisdom of 1000, he has not only guided the student to a useful article, he has also saved the student the time required to read the other 999, and arrive at judgments of their relative worth.

One of the best articles which explain the meaning of comparing numbers in the context of college rankings is the following: **Gladwell, Malcolm.** “[The order of things. What college rankings really tell us](#)” *The New Yorker* 87.1 (2011): 68-75. Downloadable copy: [Gladwell Rankings PDF](#).

In this lecture, we will read the article together. I will provide some simple and clear explanations of what is being said, so as to enable the student to read the original article. Although the article is about college rankings, it starts by illustrating the ranking problem in the context of cars. The goal of the article is to show that all rankings are deceptive – it is just one of the ways to “lie with statistics”. Even though the Car and Driver magazine comes up with a clear winner in their rankings, the winning car is NOT the best in any clear sense of the word. In fact, the question itself is meaningless. It is impossible to rank cars without consider the PURPOSE of the ranking. In this lecture, I will provide a simplified and detailed explanation of the material in the article, to enable students to read and understand the article itself.

Suppose that there are three dimensions along which cars are evaluated – Appearance, Engine, Price. Let us put aside the issue of how we come up with numbers for the subjective categories, even though this is also important. Let us suppose a panel of experts can judge, on a scale of 1 to 10, the objective rankings of cars on these three dimensions. The first concerns external appearance, style, attractiveness. The second concerns engine performance judged by many different criteria. We have omitted one of the criteria used in the Gladwell article: “the subjective feel of driving,” which has to do with how the car handles when it is driven in different situations. We have replaced this by the price, which can be evaluated objectively. Here is a set of hypothetical numbers which evaluates three cars along these three dimensions.

| Car Name | Appearance | Engine | Price |
|----------|------------|--------|-------|
| Porsche | 6 | 9 | 3 |

| | | | |
|-----------|---|---|---|
| Lotus | 8 | 7 | 6 |
| Chevrolet | 5 | 5 | 9 |

Note that high numbers mean high ranking, so the ranking of 5 given to Chevrolet means that it is the cheapest car, having the best price among the three cars being evaluated.

Note that each of the three cars is best in one of the three dimensions. Lotus is best in appearance, Porsche has the best engine, while Chevrolet has the best price. How can we find out which is the best car overall? The CORRECT answer to this question is that we CANNOT do this. The ranking between the cars depends on the PURPOSE for the evaluation – WHY are we trying to rank the cars. Without specifying a purpose, we cannot rank the cars. The standard methodology in use is deceptive – another illustration of “How to Lie with Statistics”. It assigns weights to all three factors to come up with a combined score. Let us look at how this is done. I will use C&D to be a hypothetical version of the Car & Driver magazine which is discussed in the actual article. The statements below about C&D correspond only roughly to Gladwell’s article, and are meant to simplify a more complex discussion. With this warning, we consider how C&D comes up with a ranking of cars, even though this is impossible to do without considering purpose of ranking.

C&D editors feel that what is inside the car, the engine, is the most important factor. They assign it a weight of 50%. Because they are car enthusiasts, they find that a sleek and stylish appearance is very important, and the price is not so important. So, they assign a weight of 40% to appearance, and 10% to price. Once these weights have been assigned, the score for each car can easily be calculated. Multiplying by 10 to avoid decimals, we find that, with these weights, Lotus gets 73, Porsche gets 72, while Chevrolet gets only 54. The message from this ranking is the Lotus and Porsche are close to each other and both are distinctly superior to Chevrolet. The numbers create an OBJECTIVE feel – this is not a matter of personal tastes of the C&D editors, but an objective evaluation of the characteristics of the cars.

This message is completely wrong. The rankings are created as a MIXTURE of subjective weights and objective characteristics of the cars. To bring this out, Malcolm Gladwell argues as follows. He says that Car and Drivers editors used the SAME weights for this evaluation that they do for SUV’s (Sports Utility Vehicles). Now SUV’s combine elements of practicality with a sporty feeling, but the cars being evaluated are high class luxury cars. He says that the typical buyer of sporty luxury cars is a lot more interested in the APPEARANCE of the cars, as compared to what is inside the engine. These cars are brought for show. If we change the weights to 50% on appearance and 40% on the engine, with price still at 10%, then Lotus emerges as a clear winner. The scores are now: Lotus 74, Porsche 69, and Chevrolet 54. Putting even more weight on appearance would put Lotus even further in the lead.

Next consider a buyer who has a modest income, but great love of luxury. He would be very happy to buy a sports luxury car, if only he could afford one. As long as the car is classified as a luxury car, it is all the same to him. He is maximally concerned with the price. If we put

weight of 50% on the price, and 25% each on Appearance and Engine, the Chevrolet will emerge as the winner with 70, while Lotus and Porsche lag behind with 67.5 and 52.5 respectively.

So depending on the tastes of the buyer, and the purpose for which the car is being bought, the ranking would be different. Malcolm Gladwell explains that there are two situations in which it is possible to come up with an objective ranking. One situation is when we focus on only one factor. If we look only at price, or at power of engine, or at appearance, then we can evaluate two cars X and Y and decide if X is better than Y in appearance or not.

According to Malcolm Gladwell, the second case in which objective rankings can be done is if all of the cars are similar to each other on the dimensions being ranked. He thinks that it is the diversity of the cars being ranked that leads to the sensitivity of the ranking to weights. This is a mistake. Even if the cars are similar to each other – homogenous, in Gladwell's terminology – the problem of sensitivity to weights will remain exactly the same as in a heterogeneous group. According to Gladwell, the problem arises because Car and Driver tries to cover the field and rate a very diverse group of cars. This is not true.

The source of the failure lies in the failure to specify the PURPOSE for which the ranking is being done. When we explain WHY we want to rank the cars, then we can correctly specify the weights for the different factors. The purpose is subjective – it depends on the person who is buying the car. For example, someone might allocate budget for the car to be \$20,000, and then say that he wants to get the most sporting car that he can for this price. He can then assign his personal subjective preferences for external attractiveness and engine quality to come up with a ranking. Or, he need not convert qualitative information to numbers. He could just look at cars within his budget and classify them as A,B,C – extremely attractive, attractive, and average looking – in appearance. Then, depending on his personality, he might check the engine characteristics of the A-rated cars to ensure that they are satisfactory for his purposes, and buy the most attractive one. Or he might go for a compromise between Appearance and Engine. None of these methods of choosing cars correspond to creating a ranking by numbers of the cars.

This brings us to the META-QUESTION: Why are we discussing numerical measures of car quality? This is because there has been a huge emphasis on measuring things and assigning numbers to qualitative concepts. This comes from the widely accepted Kelvin's Dictum: knowledge is only of what can be measured. The idea of "measuring" intelligence by a single number – the IQ – was invented in the 20th Century. But this is NOT a good idea. Complex multidimensional characteristics like "intelligence" cannot be reduced to a single number. In order relate knowledge to our life-experiences, we have to break knowledge out of the boxes to which it has been confined in the West. It is exploring these meta questions that leads us to the understanding of world we are living in, which has shaped our ways of thinking. It is this understanding that offers us liberation from the boxes to which education confines our thought.

2C What Do College Rankings Measure?

Part C of 2nd Lecture on Descriptive Statistics: An Islamic Approach [DSIA L02C]. We continue our study of Malcolm Gladwell's (MG) article on '[College Rankings](#)'. We will consider the questions of "How are the Numbers Computed" and "What do they Mean?"

MG starts by noting that the Purpose and Audience for College rankings have changed over time. It was initially meant as a rough guide for “consumers” (students choosing colleges). It was not imagined that Colleges would use these rankings as benchmarks of performance, proof of good management, status markers in the rivalry among colleges. It was not imagined that educational policies would be used to engineer a rise in ranking. As we have discussed, changing purposes require changes in measurements, and the rankings have NOT been changed to suit the changes in purposes, with harmful results.

Going into the methodology of the ranking itself, it is based on seven major factors:

1. Undergraduate academic reputation, 22.5 percent
2. Graduation and freshman retention rates, 20 percent
3. Faculty resources, 20 percent
4. Student selectivity, 15 percent
5. Financial resources, 10 percent
6. Graduation rate performance, 7.5 percent
7. Alumni giving, 5 percent

MG registers Two Major Complaints about the ranking process:

1. Universities being ranked are extremely diverse (heterogeneous) – How can you compare apples and oranges?
2. Each university is extremely complex and multidimensional – dozens of departments, campuses, programs – how can a SINGLE number be assigned to this?

I will explain that there are many other issues worth considering about this ranking later on. For the moment, we note that many, many questions arise from this description of the ranking process:

1. Why these SEVEN factors? Why not others? What is the basis for selection?
2. Why these weights?
3. WHAT do these factors MEASURE?
4. How are the factor scores computed?

We start the discussion by asking a simple question: Does a number actually measure what it is supposed to measure – that is, Are Numbers Accurate? To explain this, MG considers the example of the Suicide Rate. Here the Target IS Measurable – that is, there really is a NUMBER that measures the number of people who committed suicide in the past year. But no one knows what that number is. The statistics which are available are distorted by many biases. It is extremely difficult to guess the intentions of a dead person. Someone classifies a death as a suicide, and this person varies greatly by country. Depending on culture and customs, classification could be done by police officers, family, doctors. Whether or not it is reported on official statistics is again a separate matter. Because of this diversity, it would be a hopeless task to compare suicide rates across countries with any degree of confidence. INSTEAD, one should ask the PURPOSE of the comparison. If quality of life is the target, then more direct measures based on surveys of welfare may give better results.

In addition to criticisms by MG, I would like to focus on the issue that when we look at a number, it is essential to be clear about the TARGET – WHAT is that number trying to measure? So, when I am given a number measuring the Quality of a College, I must ask “What do you MEAN by Quality of College?” One way to specify this quality (and there are many other possible definitions) is to consider student learning: Student ENTERS with knowledge and skills, and EXITS with MORE knowledge, skills. The DIFFERENCE between the two is the Educational Outcome, what the college contributed to the learning process of the student. Of course, this is a Multi-Dimensional quantity – learning and skills occur on many different dimensions which are not comparable with each other. It is hard to reduce multidimensional performance to a single number unless some clear and specific purpose of education is specified. For example, if we consider how well the education provides medical skills in terms of the ability to treat patients, it might be possible to come up with a single number that aggregates the contribution of all dimensions to the single purpose. This is a complex issue, which will arise in many different contexts.

A SECOND ISSUE of importance, in terms of IMPROVING how we do statistics: From VAGUE & IMPRECISE measures of INPUTS, move to measures of OUTPUT. Stiglitz-Sen-Fitoussi recommended moving to consumption, which directly measures what human beings get, instead of production, which measures goods produced that COULD potentially get to the consumers. To illustrate this idea, consider one of the seven factors used in the rankings: Faculty Resources. According to the reasoning given for this factor, Student Engagement with faculty is an important part of the educational process. Instead of directly measuring Student Engagement (which is vague and qualitative, and hard to define and measure), we use PROXY measures, which are INPUTS that go into producing Student Engagement. These proxies are:

1. Class Size
2. Faculty Salary
3. Proportion with Ph.D.
4. Student-Faculty Ratio
5. Proportion of Full-Time Faculty.

It is true that these factors all have the POTENTIAL to create a better student educational experience. These are INPUTS into the educational process. But how effective are they? Do they actually achieve this potential? Do these factors really matter?

Suppose we specify the TARGET of our quality measure as before: “How much students “GROW” in the educational process?”. MG cites [Educational Research by Terenzini & Pascarella](#) Meta-Study of 2600 papers, which finds NO RELATIONSHIP between student engagement AND the standard list of variables used in nearly all methods for measurement of quality of colleges: educational expenditures per student, student/faculty ratios, faculty salaries, percentage of faculty with the highest degree in their field, faculty research productivity, size of the library, [or] admissions selectivity

If these INPUTS do not matter, then what DOES matter? It turns out that the key variables are the ones that are qualitative and non-measurable, or SOFT Variables: Educators engage students when they are: purposeful, inclusive, empowering, ethical, and process-oriented.

For a summary of takeaways from the Terenzini and Pascarella in-depth studies, see: [Pathways to Success: Student Engagement](#).

Focus on what can be measured takes attention away from the important qualitative factors, which often cannot be reduced to numbers. For an example of this (not discussed in the MG article) consider the question of “Do SAT scores predict academic success?”. There is a huge Controversy about the issue but the facts are clear. SAT is solidly correlated with First-Year Performance. Correlation weakens with time. BUT the effects are very small. For practical purposes, it is reasonable to conclude that we should NOT use SAT for college admissions. WHY? Again the key factors which lead to success are not measurable. Research shows that Student Characteristics strongly correlated with Success are Drive, Motivation, and Perseverance. These are character traits that are not measured by SATs. Another way to think about this question is to ask: “Can we take students with low SATs and turn them into Super Performers?” The answer is YES, and there is a lot of evidence that teachers who motivate & inspire can take students from any background and turn them into star performers.

Concluding Remarks: It is helpful to look at the bigger picture. The rise of Logical Positivism in the 20th Century led to an extreme emphasis on the observable and measurable, and complete neglect of the qualitative and unmeasurable aspects of our lives. This has led to the drive to MEASURE everything. But the Most Important Things in life are not measurable. We live our lives without measuring in numbers those things which matter the most to us – loving and being loved. This ability to deal with qualitative and unmeasurable phenomena needs to be extended to the bigger world of education and management. Even when complex, multidimensional phenomena ARE measurable, multiple measures CANNOT be reduced to one number. False philosophies lead us to PRETEND that a single number can MEASURE the quality of colleges. This type of confusion arises from a failure to think clearly about the TARGET – What is being measured and Why? To improve statistical analysis, we must learn to think clearly about the bigger questions, instead of confining attention to the numbers alone

2D Goodhart's Law

To briefly review, we note that the essence of an Islamic approach to knowledge focuses on the purpose of acquisition of knowledge. An easy way to summarize the methodology is to ask FOUR Questions:

6. Why are we doing this analysis?
7. What do the numbers mean?
8. How were the numbers computed?
9. **What is the potential IMPACT of this analysis?**

Our primary focus in this lecture will be on the fourth question. It is worth noting that these questions break the barrier between theory and application. They bring in INTENTIONS, central to an Islamic approach. They analyze relationships between external observations and internal reality. They force statisticians to enter the real world of applications, instead of staying in sterile world of theory and numbers.

As background information, it is worth noting that this idea of looking at the numbers and NOT looking at the reality behind the numbers, comes from Logical Positivism, an immensely popular theory of knowledge which emerged in the 20th Century. The central idea of this philosophy was that we and only have knowledge about the observable facts – the hidden and unobserved reality is not part of scientific knowledge. A key strategy of Logical Positivism was to replace Unobservables by Observable Manifestations. For more information about this, see [The Emergence of Logical Positivism](#).

As an example of this strategy, consider Economic Theory. There are three concepts: Welfare, Preference, Choice. Welfare refers to what is good for me (spinach), Preference refers to what I like (Ice Cream), and Choice refers to what I choose (Hamburger). Even though all three are distinct, only choice is observable. The effect of Logical Positivism on Economics, was to equate all the three:

$$\text{WELFARE} = \text{PREFERENCE} = \text{CHOICE}$$

This has BLINDED economists to the real sources of welfare, and created the confusion that everyone automatically knows what is best for her/him and DOES it. This equation leads to FREEDOM as being IDEAL. Everyone knows what is best for him and chooses it when the option is available. Behavioral economists have found that people frequently make bad choices, which are harmful for them. As a result, the NUDGE theory is based on the idea that we can guide people to make better choices in various ways which leave them free to choose, but make the best choice easier to find and make.

The numbers we measure, or Statistics, has a great IMPACT on real world. When we replace UNOBSERVABLE by OBSERVABLE & MEASURABLE Key Performance Indicators (KPIs), people also replace efforts on unobservables by efforts on observables. For example the genuine quality of research is unobservable. But we can get some idea about it by looking at the quantity (COUNT) and the reputation of the journals (Impact Factor). When this was done worldwide, it led to a massive increase in Fraud Journals, which publish articles for money and use many gimmicks to get impact factors. Instead of focusing on the unobservable reality, the focus shifted to the measurable numbers. Similar examples of how KPIs have led to attention to meaningless numbers, instead of the reality behind the numbers are available in all fields. For more illustrations, see [Beyond Numbers and Material Rewards](#).

In every area of knowledge, the positivist attempt to replace unobservables with observable and measurable quantities has led to serious problems. For example, in The MISMEASURE of MAN, Stephen Gould talks about the problem of “Reification” – replacing the abstract with the concrete. In particular, the IQ measure reduces the abstract, qualitative, multidimensional characteristic of intelligence to one number. This and other absurd ways of measuring intelligence (like shape of skull and brain size) have been taken seriously because of this tendency to replace the hidden unobservables with measurable manifestations.

With this as background, we come to the topic of our lecture. Goodhart’s Law states that “When a measure is used as a policy target, it ceases to be a good measure.” We OBSERVE something which is correlated with High Quality, and use it as a MEASURE of

Quality. Awareness that it is being used as a measure leads to change in behavior. For example, instead of trying to improve the quality of education, colleges focus on the indicators used by the reports and try to increase the numbers, leading to harmful results. For example, the college could rise in rankings by graduating everyone, regardless of academic performance. Trying to raise the ranking numbers leads to bad policies, and also DISTORTS the Numbers. When we replace Qualitative Unobserved Targets with Quantitative measures of the Target, this creates a SHIFT in the GOALS.

Next, Malcolm Gladwell examines a specific measure: The REPUTATION Score which has a 22% weight in college rankings. He asks HOW is it computed? This is done by a Survey of High School and College Officials. But do the people surveyed HAVE information required to rank colleges? Critical Question. Most people know very little about the 200+ colleges in the survey and cannot provide any useful information about this matter. Research Studies of Reputation Surveys show that they produce good results IF experts are asked about their area of expertise. Otherwise, they just replicate UNINFORMED public consensus. To prove this point, MG discusses two examples.

One is the analysis of “Best Hospitals” Rankings produced by asking doctors to rank hospitals in their area. A researcher took measures of hospital quality based on objective factors like mortality rates, type of equipment, staffing, etc., and found that objective measures of quality had zero correlation with the reputation ranking. This is simply because typical doctors do not know much about hospitals other than the ones they work in. Similarly, Lawyers were asked to rank Best Law Schools. It was found that they ranked Penn State very highly. BUT Penn State does not have a law school! This illustrates the level of ignorance about law schools together with the effect of general reputation in public perception of the school.

So how do people rank colleges when they know nothing about how to compare different colleges? They turn to public sources of information about this matter – that is, the US News and World Report College Rankings. So it is the Rankings that Drive the Reputation Score! This is an Illustration of Goodhart Law. Reputation is based on the ranking – but the ranking gives the highest weight 22% to reputation. To illustrate the difference between informed and uninformed rankings, MG mentions rankings of colleges done by corporate recruiters. Because these people take graduates and place them at jobs and follow their progress, they have knowledge about the quality of graduates being produced by different universities. Their rankings are very different from the US News and World Report Rankings, For instance, they rank Penn State at the top, even though this does not come within the top 20 in the USNWR rankings.

Goodhart's Law illustrates how our observations change the world. When we measure things, our measures acquire importance. Hidden Quality is signaled by some markers, such as Ph.D. faculty, small classes, selectivity in admissions, and high graduation rates. However, if we focus policy on improving Markers of quality, instead of on quality, this leads to major mistakes. A college could rise in rankings by hiring more Ph.D.'s, increasing selectivity in admissions, and graduating all students it admits. But none of these policies will actually have any direct impact on quality. Indeed some policies which target the indicators could actually be harmful to quality. Attempts to target the indicators will DISTORT the indicators as markers of quality. AFTER

publication count became a factor in evaluating faculty research, many faculty acquired hundreds of publications in just a few years by various shady techniques, so that publication count was no longer a good marker of quality.

What this reveals is that Ranks reflect Implicit Ideological Judgments. Factors chosen and factors excluded, as well as weights attached, represent values. However, logical positivism teaches that values are not scientific knowledge, so values are never explicitly included in analysis. Instead, they are concealed in choices of factors and weights, which creates an impression of objectivity. This is what makes statistics so dangerous – it covers ideological value judgments with a pretense of objectivity created by numbers

2E How Rankings Hide Values

Lecture ([Bit.ly/DSIA02E](https://bit.ly/DSIA02E)) is the fifth & final Part of Second Lecture on Descriptive Statistics: An Islamic Approach. The 14m video is followed by a 1600 word write-up.

How were the seven factors that enter US News & World Report (USN&WR) College Rankings chosen? We could think of a lot of important factors related to college education that are not part of the ranking factors. To answer this question, we must realize that choice of factors represents values. According to dominant popular subjective valuations, Harvard, Yale, and other elite universities have high ranks and status in public perception. This Intuitive assessment of quality comes first. When creating a ranking, we try to choose factors that will MATCH our pre-existing intuition. That is, we already know in advance of creating rankings, which colleges should come out on top. We choose factors to support this intuition. When this method for choosing factors according to pre-existing prejudices is concealed, an illusion of objectivity is created. Concealment of Values involved in the choice of weights and factors, and of the role of intuition in statistical models, is one of the important aspects of deception.

To illustrate how the choice of factors is based on values, Malcolm Gladwell notes that the Price of Education is not included, even though, for most students, this is the most important aspect of education. The authors of the survey offer no reason for this. They say that it is “Just our subjective judgment”. But a deeper reason could be that putting in the price of education could go against “intuition”. The “best” universities in the popular public image are also the most expensive ones. MG asks “What happens if we include price?” He shows that some relatively unknown universities appear in the top TEN. What does this mean? It means that some universities can provide a great education at a very low price. However, a ranking that puts a lowly university at the top would create distrust in the ranking, which may be a reason for avoiding it.

To show how arbitrary the whole business of ranking is, MG refers to an online “Rankings Game” created by Professor Slater. He has collected data on many different characteristics of Law Schools used in rankings. You can assign weights and watch the rankings change. Once you understand how the game works, you can create almost any ranking you like, and you can make any university come out on top.

Going outside the MG article, we can illustrate how public perception of power shapes the measures used for ranking. The dominant power gets to make the rules about what is measured. Before World War I, “Brittania rules the waves”, and the measure of power was defined by Sea Power, Coal Mines, and others that favored Great Britain. After the war destroyed European economies, the USA emerged as the dominant economy on the Globe. In 1934, Simon Kuznets introduces the concept of the GDP, according to which the US was the world leader. Later, when some tiny Oil Economies got higher GDP/Capita, the idea was adjusted to include a reasonable Income Distribution. That is, if a few families in the nation are very rich, that does not make the nation the richest in the world. Even later, some European economies like Switzerland overtook the USA in GNP per capita. The measure of national wealth was re-adjusted to include Infrastructure and natural resources, where the USA has a huge lead over Europe. The point is that the dominant power has the ability to dictate which factors should be used to rank nations. Currently, depending on which criteria are chosen, US or China could come out on top, reflecting the shifting balances of international power. If I choose criteria, I can make Pakistan come out on top. I would choose suicide, crime rates, Number of people who live in stable families with both parents, psychiatric patients, drugs, alcohol, percentage of the population in jail, etc. Large numbers of factors that accurately reflect human lives could be used to make Pakistan come out ahead of the USA. But what does this, or any other ranking, MEAN? This question is of central importance, and not part of conventional statistics.

MG answers that Ranks are Implicit Ideological Judgments. The choice of Factors represents Value Judgments. However, positivism teaches us that values are not scientific facts. This is why conventional statistics conceal values contained in numbers. MG provides another illustration of values by discussing two factors that are opposed to each other. One of these is Selectivity: What percentage are admitted? High Selectivity automatically leads to low ranking on Graduation Rates – This requires some explanation regarding Graduation rate. The Graduation rate is not the percentage who graduate, because that would create a bias against colleges that admit poor students. Graduation Rate is the IMPROVEMENT achieved by the university over the graduation rate that its students would have in general. To explain this better, FIRST calculate the EXPECTED rate for ADMITTED STUDENTS – Yale admits superstar students who would have 98% graduation rate. How much can Yale IMPROVE this rate? At most, it can achieve 100% which would give it 2 percentage points on this factor. As opposed to this, a college that admits poor students who have a 50% chance of graduation, and achieves 80% graduation rates, can get a score of 30%. It turns out that the low-ranked Penn State does a great job on this factor.

Given that these two factors work against each other, how should we rank Selectivity vs Graduation Rate? There is no right answer. USN&WR makes Selectivity twice as important as the Graduation Rate. This just reflects a personal preference for Yale over Penn State. Students selecting colleges may prefer low selectivity, as it maximizes their chances of getting into the college. Governments wanting to fund education may prefer large public universities which admit everybody and do the best possible job on the worst students. The choice of weights, and the rankings, depend on the purpose for which it is done.

Rankings are not objective; rather they are ideologies masquerading as objective numbers. Impact of Ideology. What is the IMPACT of these ideologies? The college rankings are no longer a harmless game that we play with numbers. These rankings have a massive impact on public mindset, funding, choices made by students, professors, salaries. It is a matter of great importance that these rankings ignore the price. This means the colleges do not have an incentive to provide good education at the lowest possible price, because this would not affect the ranking. The bias in favor of the wealthy is shown by the fact that the top 20 schools are always the private elite class schools. Most of the ranking factors relate strongly to WEALTH. Heavy endowments of private universities make them impossible to compete with. Even though Penn State is the Most Popular university – 115,000 applications – but there is no way they can get into the top 20, without Billions of dollars.

Malcolm Gladwell concludes that Rankings reflect the mindset of Ranker. He gives the example of professor Huntington who took a survey of his colleague asking them to rank civilizations around the globe. There is no surprise that the survey ranked USA and UK at the top, given that all respondents were familiar with these civilizations, and had no knowledge of others.

Our own concluding remarks for this second lecture are as follows. There is a famous saying that “Statistics are the eyes of the State”. Factors that are measured get attention, while aspects of society that are not measured tend not to receive attention in public policy. In particular, the most important numbers are the GDP, where there is a concerted effort by Ministries around the globe to improve rankings in GDP. But the target of the efforts is the NUMBER and not the reality behind the number. The Bureau of Statistics can do many kinds of manipulation to increase GDP and increase growth rates, without having any effect on the lives of the people. Focusing on NUMBERS leads to harmful policies, while focusing on the REALITY that numbers are meant to measure would improve policies. This point is made forcefully in a book by Stiglitz with the title: “Mismeasuring our Lives: Why GNP does not add up.”. He shows how essential aspects of our lives are not measured in GNP, and this leads to very poor economic policies which cause a lot of harm in the dimensions which are not measured, while improving the measured dimensions.

This problem does not relate to GDP or College Rankings alone. Rather, we use numbers to measure performance in many dimensions, and these same problems arise in nearly all the rankings that we use. An Islamic Approach to statistics requires us to do two things in this context. One is to make the value judgments in choice of factors and weights explicit and aligned with Islamic values. The second is to focus on the Reality behind the numbers, and not the numbers by themselves.

2F The Why of Global Corruption Rankings

In this lecture, we will examine global corruption rankings in light of The Four Questions which are central to the Islamic Approach:

1. WHY are we doing corruption rankings of countries?
2. What do the numbers mean?

3. How are they calculated?
4. What is the impact of the creation of these corruption rankings?

This lecture is based on Zaman, Asad and Rahim, Faiz, [Corruption: Measuring the Unmeasurable](#) Humanomics, Vol. 25, No. 2, pp. 117-126, June 2009. <https://ssrn.com/abstract=1309131>

We start by thinking about “*Why do we assign NUMBERS to corruption?*” After all, it is a qualitative condition of the heart, not subject to measurement. There is a long and complicated story that led to the attempts to measure the unmeasurable, which we summarize very briefly, to explain this:

1. A battle between Science and Religion fought in Europe led to the rejection of Christianity, and the acceptance of Science as the new religion of the West.
2. It became widely believed that Science is the only source of reliable knowledge. This led to a rejection of heart, emotions, subjectivity. Logical positivists introduced the Fact/Value distinction, and said science was about facts, while values were not scientific.
3. Advances in Physics were tied to accuracy of measurement. This led to the misconception known as Lord Kelvin’s Dictum: If you cannot measure it, you don’t know what you are talking about. Numbers = Knowledge. See [Lord Kelvin’s Blunder](#).
4. In the early 20th Century, Social Sciences were constructed by the application of the Scientific Method. But the methodology of science was VASTLY misunderstood by Logical Positivists and it was this misunderstanding of science that was used to create a methodology for economics, econometrics, and statistics.
5. These developments, where knowledge required measurement, led to attempts to Measure the UnMeasurable throughout the social sciences.

Can Corruption be Measured? Obviously, the internal Qualitative, corruption of hearts, cannot be measured in numbers. BUT External Manifestation, like Bribes, can be measured. It is worthwhile to define BRIBE as the Use of Money for Persuasion towards a personally profitable agenda at social cost.

Even if we confine attention to bribes, corruption is multidimensional and cannot be reduced to a single number. To see this, compare two countries. A has 100 corrupt transactions of \$ 1M each. B has 1M corrupt transactions of \$100. WHICH country is MORE corrupt, A or B? There is NO OBJECTIVE answer to this question. To answer, we need to specify the purpose of making the comparison.

There are situations when it becomes necessary to try to measure the unmeasurable. In such situations, the following *Rules for Measurement* are worth remembering.

The simplest case occurs when the Target is ONE-DIMENSIONAL and Quantitative. IN this case ONLY, objective measurement is possible. Much more often we have the case of a qualitative and multidimensional phenomenon. In this case, we should explain clearly the subjective choices required to convert qualitative & multidimensional measures into a single

number. If we consider a range of options, and also the purpose for which different USERS may find it useful, we will find different numbers for different users. This would be helpful to dispel the image of objectivity created by statistics.

Now, we come to the topic of the lecture. *How is the CPI (The Corruption Perception Index) computed by Transparency International?* To the best of our knowledge, they poll a group of wealthy businessmen of unknown identity, and ask them to rank countries from 1 to 10 in terms of their perceptions of corruption in a given country. High numbers are high honesty and integrity, while low numbers correspond to high corruption.

As discussed earlier in "[What do College Rankings Measure?](#)", the crucial question is: "How much KNOWLEDGE do they have of global corruption, and of RELATIVE corruption?" Uninformed rankings just report the prejudices of the people who are doing the ranking. There are many reasons to suspect that these rankings are done by foreigners with little knowledge of local culture. Furthermore, it is likely that these businessmen make brief visits to get big jobs done in the fastest way made possible by wealth – they look for corrupt counterparts to avoid the regular process. In any case, it is likely that the perceptions just reflect the prejudices of those doing the ranking, rather than any characteristic of the country.

What do the CPI numbers mean? Statistical analysis in [Zaman and Rahim \(2009\)](#) shows that the CPI has a 98% correlation with log (GNP per capita). In other words, Integrity and Honesty are just other names for wealth. This likely reflects the prejudice of the wealthy. In real life, we see that More wealth = more greed & corruption. The Quran also mentions how excess wealth leads to corruption. Remember that Corruption is a two-party transaction. The poor accept money to do favors for the rich – the poor get the blame and coverage, while the wealthy escape attracting attention.

If we use the definition of bribe given above, LOBBYING in the USA is easily seen to be bribery: the use of money to pursue narrow group interests while inflicting huge costs on society. The Global Financial Crisis is one example of how rich financiers got trillions of dollars in bailouts, at the expense of poor mortgagors made homeless by the millions. Another egregious example is the [Medicare Prescription Drug Bill](#) passed in 2003 using dirty tactics by [Senator Tauzin](#) on behalf of Big Pharma. The bill ensures that the Pharma industry can charge whatever price they like for sales to medicare. The government cannot negotiate, and they cannot import cheaper alternatives from Canada. The bill has been called an \$80 billion per annum give-away to the Pharmaceutical Industry. (Despite a campaign promise to do so, Obama was unable to get this bill repealed due to the powerful Big Pharma Lobby.) Afterward, Senator Tauzin left Congress to take up a \$2 million consultancy, and also received more than \$11 million in cash rewards from grateful Big Pharma. But while all of this is documented, none of this is counted as corruption!

Our Islamic approach requires us to dig deeper into the historical context and background of the numbers we analyze. *Why CPI was developed?* The answer is somewhat complex. In the Post WW2 era, there was a competition between Capitalist & Communist models of development. The World Bank offered the Structural Adjustment Program, as a roadmap for

development. There is not a single instance of success – no country became developed by following World Bank advice, but many countries, like East Asia, did industrialize by REJECTING World Bank advice (see [Choosing our own pathways to progress](#)). This failure of the capitalist model is widely documented and acknowledged by all parties. In order to maintain credibility, it was necessary for the World Bank to find some scapegoat to blame for the failure of the capitalist model for development. This was done by putting the blame on the poor countries for their own failure – a standard illustration of blaming the victim. It was not bad models created by the World Bank which led to failure, but bad governance and corruption in the poor countries which caused the failure. For more details see the article on [Michael Foucault Power/Knowledge](#) which explains how the powerful shape knowledge for their benefit.

The fourth question is: “*What is the IMPACT of CPI?*.. We could imagine that, theoretically, a country with a high CPI will make efforts to improve in terms of governance and corruption. Practically, it has the opposite effect. Solid research establishes that my behavior is affected by my PERCEPTION of social norms (and not by the REALITY). So If PERCEPTION of high corruption is created, people will act in more corrupt ways. If PERCEPTION of justice and low corruption is created, people act honestly. This means that the strategy of moving towards greater integrity and honesty is the opposite of the one currently being followed all over the developing world. Institutions like NAB and Anti-Corruption drives highlight corruption and cause it to spread. Instead, an effective strategy would highlight honesty and integrity. If a country has 99 incidents of corruption and one of integrity, publicity for the solitary good incident would create an impression of honesty and help to spread it. Thus, attempts to measure corruption via CPI are likely to be counter-productive rather than helpful in combating corruption.

Conclusions: The Colonization of the Globe was justified by Racist arguments. The white man was infinitely superior to all other races, and had the right to rule the world. The [colonization process was so extremely brutal and ruthless](#), that records have been suppressed from history and memory. Today, this process of colonization continues by financial means. Poor countries make billions of dollars of interest payments to the rich. Justification for this exploitation of the poor by the rich and powerful is still needed. This justification is created by the CPI as well as many types of economic theories of development.

Chapter Summary/Key Takeaways

Insert content here...

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

- If you want to help your child get ahead in life, learn their strengths and weaknesses.
- Point 2 from your text...
- Point 3 from your text...
- etc.

In the next chapter, you will learn...

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter, link the next chapter's information with what has already been learned.

Chapter 3: Life Expectancies

This is an Introduction to the Basic Concepts regarding Life Expectancy which will be the topic of this lecture.

3A: Many Kinds of Numbers

It is widely believed that “you can’t argue with the numbers”. Contrary to popular belief, there are many kinds of Numbers:

1. **Qualitative and Unmeasurable things:** Corruption, Love, Intelligence, Courage, Sympathy, etc. One can measure manifestations, like the number of articles for quality of research. Such Numbers can be harmful by directing attention away from reality towards the manifestation.
2. **Quantitative, Multidimensional:** Material Wealth, Body Size – these are things that can be measured, but there are many different measurements required. When multiple numerical measurements are reduced to a single number, this always involves subjective choices of factors and weights. In such cases, numbers are mixtures of the subjective and the objective, combinations of facts and values. These can be useful if the purpose of measurement is clear, and values are explicit.
3. **One Dimensional Quantitative:** These numbers are of special importance. They provide objective measures of features of external reality.

Essential Aspect of Islamic Methodology

When we ask the four questions about why, how, meaning, and impact of numbers, it is impossible to be talking about numbers in the abstract. These questions can ONLY be answered when we are talking about numbers used in a particular real-world context. Thus, we reject the separation of theory and practice, and argue that theory can only be understood within the context of real-world applications.

In this lecture, we will discuss numbers used to measure Life Expectancy (LE). LE is a useful measure of a real and important aspect of our personal lives. The amount of time that remains for us to live is of vital importance to us in making our plans for the future, and for determining how we live in the now.

Central Questions

Why are we studying Life Expectancy? Unlike GNP and other measures which are mixtures of subjective and objective, LE provides an objective ranking of countries with respect to one dimension of health.

What do the Life Expectancy numbers mean? Roughly speaking, this is the average amount of time that a person can expect to live.

How are LE numbers computed? This aspect will be discussed in detail later.

What kinds of impact can studying these numbers have on the real world? Increases in LE often correspond to increases in health and nutrition for the general population. To understand how this matters, we refer to a recently published article: "[Life Expectancy and Mortality Rates in the United States, 1959-2017](#)" by Woolf & Schoomaker. This article states that, after increasing continuously from 1959 to 2014, from 69.9 to 80, LE has declined for three straight years in a row! WHY has life expectancy started DECREASING in the USA, contrary to global patterns? The study traces the source of the decline to increases in middle-age mortality. Middle-aged people in the USA have increasing mortality rates due to drinking, drugs, and depression leading to suicides. Since Life Expectancy is increasing around the globe and also increasing in OTHER age groups in the USA, these statistics signal something extremely wrong with LIFE in middle age in USA. The statistics point to an aspect of the real world that is worth studying further in greater detail.

Using Life Expectancy as a proxy for Health

While assessing "health" is difficult because it is qualitative, LE provides a useful quantitative numerical proxy. Using this proxy, we can ask questions like the following:

1. Which countries have had rapid increases?
2. Which countries are performing well?
3. Which are doing poorly, in terms of providing satisfactory health and nutrition to their people?

Life Expectancy statistics are available and provide information on these issues.

What does life expectancy MEAN?

LE is defined as "The average age which a newborn baby EXPECTS to live". This definition does not make sense – everyone will live to some particular FIXED age – how can we take an AVERAGE over this?

SOLUTION: Take 1000 people. For each person calculate AGE, to get A(1), A(2), ..., A(1000). Now take the average. Sum of all is TOTAL age lived by the entire population of newborns, divide by the total number of newborns. This is the average age of all of the 1000 people born today. We explain how this is done by a *Hypothetical Example of Calculation of Life Expectancy*:

Of 1000 people 850 survive to 10 years, 150 die in the 0-10 category. For these 150, the combined total age is $150 \times 5 = 750$ years. This is under the rough assumption that all of them lived for exactly 5 years, the midpoint of the category of 0-10 years.

From the 850 people alive in the 10-20 category, 500 survive to 20 years. 350 died at average age of 15. The total age of these 350 is $350 \times 15 = 5250$ person-years

Of the 500 people left alive in the 20-30 category, only 100 survive to 30 years. 400 died at average age of 25. The total age of people dying in this category is $400 \times 25 = 10,000$ person-years.

Finally, of the 100 people left alive in the 30-40 age category, all 100 of them die within this range. This gives us 100 people with average age of 35, for a total age of $100 \times 35 = 3500$ person-years

TOTAL Life in person-years of all 1000 is $750+5250+10,000+3500 = 19500$. Divide this by 1000 to get Life expectancy of 19.5. This is the average age of the entire population of 1000 people. There is another statistic that is also useful for measuring the “average”. This is called the MEDIAN, or the Half-Life in the context of atomic particles. This is explained below

MEDIAN Life Expectancy, also called Half-Life

In the above example, 150 died at <10, 350 died at <20, 400 at <30, and 100 at <40. So half of the population 500 died at <20. The other half, 500 died at >20. Thus age 20 divides the population in half. Half die at age less than 20, and exactly half are alive beyond the age of 20. Thus the MEDIAN Life expectancy, or the half-life, of this population of 1000 people, is 20. Another way to say this is to ask: "what is the life expectancy of a person chosen at random from this population of 1000 people?". For the randomly chosen person, there is a 50% chance of death before 20 and a 50% chance of death after 20. This is another way to define the median life expectancy, or the half-life. Note that this is close to 19.5 AVERAGE, but somewhat different in meaning and in method of calculation.

How to CALCULATE Life Expectancy?

Life expectancy requires information that we don't have. For a NEWBORN batch of 1000 people, we need to PREDICT what % will die before 10, what % before 20, what % before 30, etc. BUT these numbers are unknown. We don't know how many of the newborns will die before their 10th birthday. The standard method to make these predictions is to use the CURRENT mortality rate. That is, look at the percentage of the 0-10 population which died in the CURRENT year, and use that to forecast the mortality of the newborns. Similarly, we can find the mortality rate of the current population of 10-20 year olds, by looking at what percentage of people in this age range died in the current year. Then we can use this percentage as a forecast for what will happen to newborns who reach the age category of 10-20 after 10 years. A “*mortality table*” gives us percentages of deaths in each age category:

Look at 2019 data, How many people aged 10-19 died? What is the total population in the age range 10-19. This ratio of deaths to total population is the MORTALITY rate for 10-19.

Life Expectancy is calculated by using current mortality rates to predict future mortality.

Conclusions

The standard method for ranking countries uses GNP per capita. However, this number is a mixture of subjective values which lead to choices of factors and weights which go into the computation of this number. The WDI tables give at least 7 variants of this and many more could be devised. These numbers are arbitrary because they are based on a mixture of facts about the countries, and of the values which go into selecting the facts to highlight, and the weights used to prioritize them.

In contrast, Life Expectancy comparisons are objective indicators of a fact about external reality. This is because we are computing a ONE-DIMENSIONAL measure of a quantitative and numerical fact about the country. Anyone who tries to compute this number, using any reasonable method, will come up with similar numbers. This number is a feature of the country, and not a projection of my ideas about how to evaluate countries. At the same time, life expectancy is not a perfect measure of the target number we are trying to estimate – the average life of newborns.

This is because Life Expectancy depends on approximating future mortality rates by current mortality rates – this approximation may fail to be valid for many possible reasons – future death rates are unknown. LE is a reasonable GUESS about life expectancy.

3B Computation of Life Expectancy from Mortality Tables

In this Part B of 3rd Lecture on Descriptive Statistics: An Islamic Approach, we explain in detail how life expectancies are computed using a hypothetical example.

A Mortality Table

To compute life expectancies, we start by constructing a mortality table. This is given below:

| Age Group | Number in 2017 | Mortality | Rates |
|-----------|----------------|-----------|--------|
| 0-9.999 | 30M | 6M | 20% |
| 10-19.999 | 40M | 10M | 25% |
| 20-29.999 | 25M | 15M | 60% |
| 30-39.999 | 15M | 10M | 66.67% |

| | | | |
|---------------|----|----|------|
| 40- 49.999 | 5M | 5M | 100% |
|---------------|----|----|------|

In each of the 5 categories of age groups, we ASSUME population size as listed in the table – M stands for Million. So there are 30,40,25,15, and 5 Million people in age categories of 0-10, 10-20, 20-30, 30-40, and 40-50 respectively. By collecting data on deaths by age category in 2017, we can find out the number of deaths in each age category. This number (hypothetical) is listed in the third column. The last column gives the mortality rate in each age category, which is just the percentage of deaths within that age category. This is called a mortality table.

From Mortality to Life Expectancy

| Age Group | Live at Beginning | Mortality Rate | Died during period | Total Lifespan |
|-----------|-------------------|----------------|--------------------|----------------|
| 0-9.999 | 1000 | 20% | 200 | 1000 |
| 10-19.999 | 800 | 25% | 200 | 3000 |
| 20-29.999 | 600 | 60% | 360 | 9000 |
| 30-39.999 | 240 | 67% | 160 | 5600 |
| 40-49.999 | 80 | 100% | 80 | 3600 |
| | | | SUM= | 18,600 |

To compute Life Expectancy at birth, we do a thought experiment. Imagine a batch of 1000 people who are all born on Jan 1, 2018 – what will their average age be? This is the life expectancy. We use the mortality table to GUESS that 20% of these people will die within the 0-10 category – that is, 200 people will die before reaching their 10th birthday. Now we ask: how long do these people live, cumulatively? We make the assumption that these 200 deaths are **EQUALLY** distributed over the 10 years, so that there are 20 deaths each year. With this assumption, the average age of all 200 people will be 5 years, the midpoint of the 0-10 age category. This means that the total lifespan of these 200 people who died will be $200 \times 5 = 1000$ person-years. This gives the numbers in the first line of the table.

For the second line, we start with the 800 survivors of our batch of 1000 newborns, all of whom have birthday Jan 1, 2018. It is now Jan 1, 2018, and ten years have passed. 200 people in this batch died, and 800 are still alive on this date. How many will survive from 2018 to 2028, and come into the 20-30 age category? The mortality rate for 10-20 year olds that we see in 2017 is 25%, and this is our best guess for what might happen in the future. Applying this rate, we guess that 25% of these 800 people will die, leading to 200 deaths. How long did these 200 people live, cumulatively? Assuming equal distribution of deaths over the ten years, the average age of this group would be 15 years, the midpoint of the 10-20 category. This gives us 3000

person-years which is 15 years times 200 persons, as the total lifespan of all 200. This gives us the numbers in the second line of the table, for the age group 10-20.

We can go on to the third line and complete it in the same way. 600 people will be alive in Jan 1, 2038 and enter into the 20-30 age category. The mortality rate in this category in 2017 is 60%. Applying this rate to the future, we assume that 60% of these 600 people will die before Jan 1, 2048, while in the 20-30 age category. The cumulative age of the 360 people who will die is $360 \times 25 = 9000$ person-years.

Continuing in this way, 240 people will be alive on Jan 1, 2048, and will make it into the 30-40 age category. A mortality rate of 66.67% for this age category in 2017 leads us to estimate that 160 of these people will die before reaching Jan 1, 2058. The average age of these 160 will be 35, so the cumulative life experience of this batch will be $160 \times 35 = 5600$. Only 80 people will make it to Jan 1, 2068, entering the 40-50 age category. We assume 100% mortality in this age group, so all of them will die. The cumulative life experience of this last batch will be $80 \times 45 = 3600$.

Now the life expectancy can be computed by adding up all of these lifespans of all of the 1000 people. This sums to a total of 18,600 person-years. If each of the 1000 people lived exactly 18.6 years, then the sum total would also be 18,600, so 18.6 is the average age of these 1000 people. This is the life expectancy.

The Median or Half-Life

Instead of the average lifespan, we can also use the median lifespan as a measure of the life expectancy. The median lifespan for these 1000 people is the date on which 500 people die, while 500 people remain alive. This can easily be computed from the previous table. In the first two age categories, from 0-10 and 10-20, we have 400 deaths. In the next category of 20-30, we have 360 deaths, so that the cumulative deaths at the end of this period are 760. Obviously, the 500th death will occur within this age category. We know that 600 people entered this age category of 20-30 on Jan 1, 2038. We know that 360 of them will die in the next decade, between Jan 1, 2038 and Jan 1, 2048. The median age is the date on which the 500th person dies. 400 people have already died earlier in the previous age categories. So the 500th death will be the 100th death in the current batch. A total of 360 deaths will take place over the ten years. So the 100th death will take place at a time which 100/360 proportion of the 10 year period, assuming deaths are evenly spaced over this decade. This leads to $2.777 \text{ yrs} = (100/360) * 10 \text{ yrs}$. Thus the median lifespan the 22.777 years, which is the predicted age of the 500th person to die in this batch.

Refinements: Equal Deaths vs Equal Probabilities

The above calculations are just a rough first pass attempt at explaining the details of life expectancy calculations, according to the “period life expectancy” method. There are other methods, such as the “cohort” method, which can also be used. A discussion of life expectancy numbers and what they mean is given in “[Life Expectancy – What does this actually mean?](#)” by [Esteban Ortiz-Ospina](#) on World Bank website, on August 28, 2017. Regardless of

how we make the computations, life expectancy depends on making extrapolations about the future based on what we have observed in the past. The future is inherently uncertain, and hence these projections never be made with great accuracy. It is best to accept the uncertainty, and use simple methods, rather than use fancy methods to create an illusion of sophistication and accuracy which cannot be achieved when guessing about the future.

There is one refinement which can sometimes be important, and is worth explaining and clarifying. Our projections, as calculated above, are based on the assumption that if there are 200 deaths over 10 years, then these are equally distributed over the 10 years, so that there are 20 deaths each year. For the 0-10 age category, we can detail this as follows:

| Age | Survivors | Deaths | Rate |
|------|-----------|--------|-------|
| 0-1 | 1000 | 20 | 2.00% |
| 1-2 | 980 | 20 | 2.04% |
| 2-3 | 960 | 20 | 2.08% |
| 3-4 | 940 | 20 | 2.13% |
| 4-5 | 920 | 20 | 2.17% |
| 5-6 | 900 | 20 | 2.22% |
| 6-7 | 880 | 20 | 2.27% |
| 7-8 | 860 | 20 | 2.33% |
| 8-9 | 840 | 20 | 2.38% |
| 9-10 | 820 | 20 | 2.44% |
| | | 200 | |

Equalizing the number of deaths leads to a rising probability of death. Alternatively, it would be possible to equalize the probability of death, which would make the number of deaths decrease. It is not possible to prefer one method or the other on theoretical grounds. The best solution is to go to the data to actually get the death rates by 1 year categories and avoid theoretical approximations. However, the reason for mentioning this is that it sometimes the highest mortality occurs in the 0-1 age category. Once people survive to one year, then the mortality goes down dramatically. In such cases, the assumption of equal deaths, or equal probability of death, can both be misleading.

Our goal in this lecture is not to make students experts in demography, and in learning the methods of computation of life expectancy. Rather, we are providing enough details to ensure that the students are comfortable with the concept, and understand how it is calculated and what

it means. It is important to note that even though the life expectancy is a guess about the future, it is based solidly on current data. This means that it reflects the realities of mortality today, and gives us useful information about population health today, wrapped into one number.

3C Analyzing World Bank Data on Life Expectancies

Part C of Lecture 3 on Descriptive Statistics: An Islamic Approach, starts the analysis of the WDI (World Development Indicators) data set of the World Bank, which covers 190 countries from 1960 to 2018. The goals of our analysis are different from the goals of conventional analysis. Sir Ronald Fisher, also known as the father of statistics, defined statistics to be the reduction of data. In contrast, an Islamic approach seeks to produce useful knowledge, instead of playing games with numbers. Thus, we aim to use these numbers to learn about the real world. For this purpose, it is important to choose numbers that convey information about the real world. Life Expectancy is one such number, which is a piece of objective information about living conditions in any given country. This is different from data like GNP per capita, which is a mixture of subjective opinions about what is wealth, and what are relevant and irrelevant factors to be considered for this purpose, with genuine data about the real world. Because of this mixture of facts and values, many numbers reflect the mindset of the creator of the number, rather than a fact about external reality. For conventional statistics, it does not matter what kind of numbers we have, and how they are related to external reality. Summarizing the numbers involves the same set of operations. But for an Islamic approach, there is a huge difference, because numbers that reflect subjective mindsets will not help us to learn about the external reality. Video linked below provides the first steps towards learning about this life expectancy data set, and what it tells us about the real world.

====WRITEUP (also: attach spreadsheets in 3D)

One of the important lessons conveyed in this lecture is the importance of benchmarks for comparison. If we went to assess the performance of any country, we have to COMPARE it with something else — that something else is a “benchmark” for performance. We start by creating three benchmarks. the Minimum, the Maximum, and the Median. By using these three numbers we can look at any country and place it in the bottom half or the top half of the countries with respect to Life Expectancy. We can also judge how close it is to the top and bottom rankings to get an idea of its position among all the 190 countries for which the data has been tabulated. Such comparisons allow us to judge performance, as to whether it has been good, bad, or average. Whenever we say the X is good or bad, we are making a COMPARISON of the performance of X with something else. For clarity, it is essential to state what X is being compared with, to evaluate its performance. Choosing benchmarks, and justifying them is one of the essential aspects of the rhetoric of statistics, the methodology by which numbers are used to persuade.

3D Histograms for the World Bank Life Expectancy Data

Some Historical Background: How do we analyze a large data set? Annual LE data for 190 countries 1960-2017 consists of 57×190 is more than 10,000 points of data?

Classical Answer: Find an approximate theoretical model for it. Fit data to model, then analyze properties of the model. This “reduces” the data by replacing it with a theoretical model with a few parameters, and makes it easy to understand and analyze. But WHY use this method, which now appears natural to statisticians? BECAUSE Computational Capabilities to do the right kind of analysis DID NOT EXIST!!

Currently, Statistics is caught in a theory trap – Computational Capabilities required for good data analysis have come into existence – though only recently. Pedagogy has NOT caught up to this. We are still teaching statistics as if computers do not exist.

What is the GOAL of statistical data analysis?

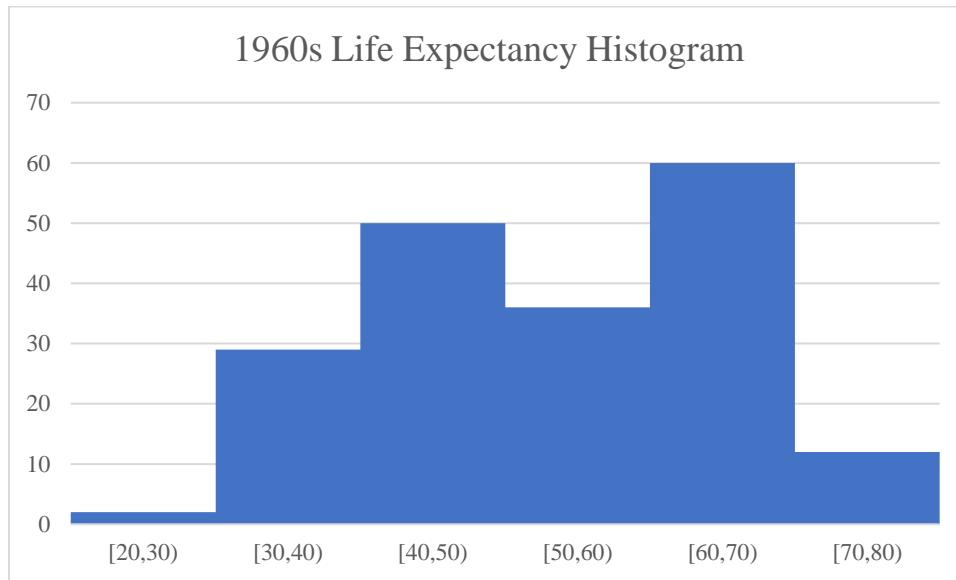
1. FIRST: to UNDERSTAND the data set – what do the 190×57 LE numbers tell us?
2. SECOND: to trace implications of this message for reality

PROBLEM – Our minds are NOT equipped to understand raw data, a table of 190×57 numbers. We cannot see patterns in this data. We cannot even find the minimum, maximum, or compare and evaluate countries, using the TABLE directly.

SOLUTION – FIND a way to represent the data that makes SENSE to our MIND. We are looking for a GUI – a Graphical User Interface – to access the data. Prior to Computers, theoretical models of the data were the ONLY possible approach to reduce the data. A HUGE amount of work exists on HOW to fit theoretical models to data. Conventional statistics is ALL ABOUT theoretical models for data. Our “Radical Approach” is based on CHANGING the goals – instead of creating a theoretical model to represent the data in a form that is “easier” to understand, we try to DIRECTLY understand the data ITSELF! This lecture will be about one of the best tools which permits us to look directly at the data: the HISTOGRAM.

How do we make a histogram? Categorize the data, and look at numbers in each category. For example, suppose we have demographic data on a population, classified by age. We can count the number of people in the following age categories: 0-10 children, 11-20 teenagers, 20-40 Younger workforce, 40-60 older workforce, 60+ retired. Each of these categories is called a “bin”. Many other ways to categorize the data are possible. A PICTURE of the data which splits data into categories, and COUNTS the number of points in each category, is called a HISTOGRAM. In this lecture, we will MAKE, interpret, and analyze, histograms of the life expectancy data.

The MAKING part is easy because EXCEL now has a built-in Histogram Graph type. This was not available in previous versions, and it was quite clumsy and difficult to make histograms before 2016. Just highlight a column (or row) of numbers, and click on the Histogram graph type to make a histogram. Below is a Histogram of Life Expectancy in 1960. This puts data for 190 countries into 6 bins from [20,30) in ten-year categories going up to [70,80). As a notational convention “[” means inclusion, while “)” means exclusion; [20,30) means all ages including 20, and going up to 30, but excluding 30. The histogram is given below. The entire World Bank WDI data set on Life Expectancies is available via shortlink: <http://bit.ly/rsraC3LE>. Students can examine the spreadsheet and use EXCEL to replicate the histogram above. Instructions are provided on the EXCEL spreadsheet.



What do we learn from this histogram? A technical term is useful for proceeding with this analysis. The MODAL BIN (short: MODE) is the one which contains the largest number of data points. From the picture above, it is clear that the MODE occurs in the range [60,70). There are 60 countries in this age category, more than any other category. This means that the largest number of countries (60) had life-expectancies in the range 60 to 70 years in 1960. There are a few countries (12) which have higher life expectancy in the range 70-80. There are two countries which have life-expectancies below 30. The histogram does not tell us which country lies in which bin, but we can learn that by going back to the original data. The bottom 6 countries are:

| | |
|--------------|--------|
| Mali | 28.199 |
| Yemen, Rep. | 29.919 |
| Sierra Leone | 31.566 |
| South Sudan | 31.697 |
| Gambia, The | 32.054 |
| Afghanistan | 32.446 |

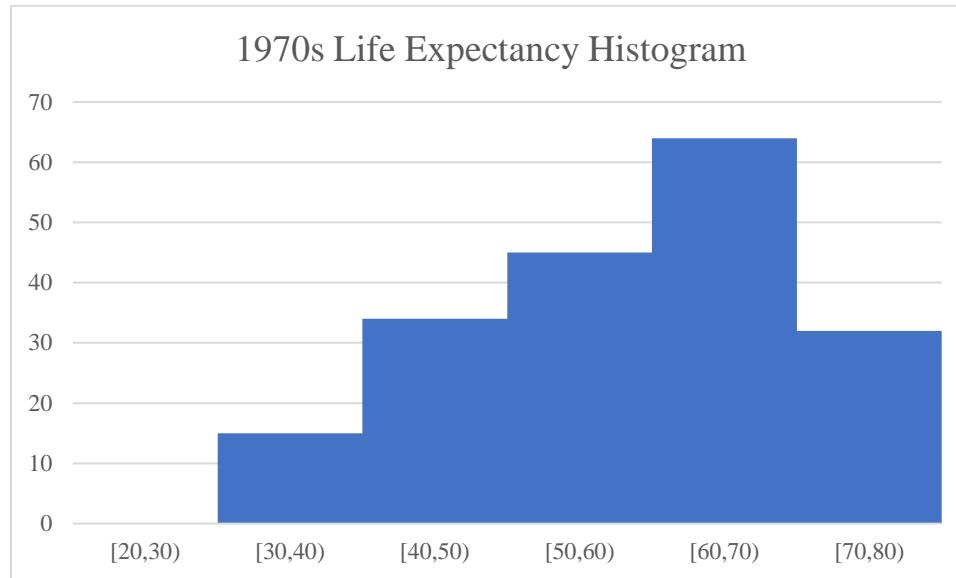
The way we choose to create the bins is arbitrary, and this makes it possible to distort the data, and make some wrong arguments. For example, the bottom bin goes from 20 to 30 and contains

only two countries, Mali and Yemen, with life expectancies of 28.1 and 29.9 respectively. From the histogram, one could get the impression that these two countries have exceptionally low life expectancies. Even though these countries are at the bottom, looking at nearby countries, we see that there is no clear cutoff between these countries and those slightly above them. In fact, there are about 12 countries in the range of 28 to 35. These countries have the lowest life expectancies in the data set, but they are all close to each other, and none could be termed exceptional within this group.

The global mode is the bin with the largest number of data points within it. With bin sizes as specified, we see that the category [60,70) is the global mode. The concept of a “local” mode is also important in understanding data sets. A local mode is a bin which has more data than its neighboring bins on both sides. In the above data set, the bin [40,60) is a local mode with 50 countries, more than the neighboring bins of [30,40) and [60,70). When a histogram has two modes, it is called bimodal. A bimodal data set SUGGESTS a hypothesis about the real world: there are two different types of countries, those with a low Life expectancy and those with high Life Expectancy. This is just guess based on some ideas about how data tends to be distributed, but it is also supported by the well-known fact that there is huge divide between countries in terms of wealth. It stands to reason that life expectancies would be high in the wealthy countries and low in the poorer countries. However, this natural hypothesis leads to puzzle. The wealthier countries are few in number while the poor countries are much more numerous. If wealth and life expectancy were very highly correlated, we would expect to see a high mode with low life expectancy, and a smaller mode with high life expectancy. In fact the graph shows the opposite. To investigate this discrepancy, we would need to look at the countries in the two modes, and try to find out the factors which differentiate between them. We would look for those factors which are relevant to increased/decreased life expectancy.

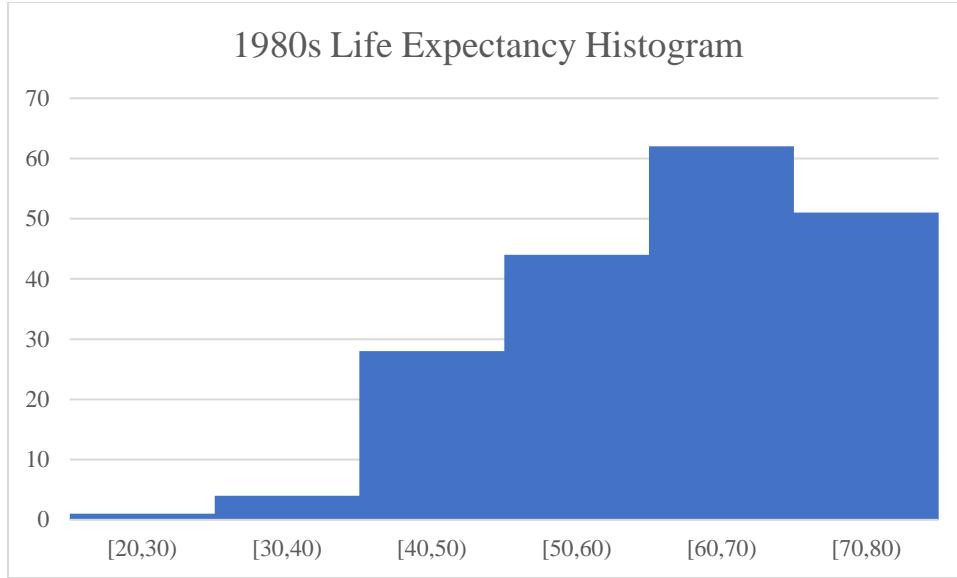
This brings us to a central methodological difference between “real” statistics, and the conventional approach. Real statistics involves going back and forth between data analysis and looking at qualitative aspects of the real world. Data analysis above suggests that there are significant differences in life expectancies between countries. But the data cannot tell us WHY? To learn why, we must go beyond the data set, and try to understand what features of the real world lead to this difference. Once we have a handle on the relevant factors, then we can again try to measure and quantify these factors, to assess how important they are in explaining differences in life expectancies. Success or failure in this process will lead us to a deeper understanding of the causes of high life expectancies, and perhaps provide us with guidance on policies we could undertake to improve life expectancy. It is essential to realize that crunching numbers without analyzing the real world cannot lead to similar depth of understanding. As Freedman said, understanding causes requires expending “shoe leather” – we have to walk around in the real world, to understand the causes of higher life expectancy. Whereas conventional statistics tell us that we study data to arrive at conclusions about the real world, “real” statistics methodology tells us that we study data to generate conjectures about the real world. These conjectures must be assessed by direct study of relevant real-world issues.

We can study changes in country average life expectancies through time by looking at how the histogram changes. For this purpose, we redo the LE histogram for 1970:



This histogram shows substantial progress in life expectancies from 1960 to 1970. The number of countries with LE less than 40 has gone from 31 to 14. The second mode at [40,50) has disappeared; from 50 countries this category has shrunk to 33. Reductions in the bottom three categories are matched by increases in the top three categories. Countries are rapidly moving up the scale of life expectancy. Whatever the factors which lead to this substantial increase life expectancy, they are spread broadly over the world, and not concentrated in the small number of extremely wealthy countries. One would guess that easy availability of cheap antibiotics across the globe may be an important part of the explanation. To verify this, one would need to gather data about how use of antibiotics has spread over time, and whether change is parallel to the increases we see in life expectancy. It worth noting the machine learning and mechanical analysis of big data cannot generate hypotheses about the factors which are potentially relevant and worth studying – that requires deeper understanding of the real world. Once the data has been collected – based on intuition and knowledge – then statistical techniques come into play in verifying our hypotheses about the causes for the changes we observe. Statistics cannot tell us which data we should collect.

Finally, we look at the histogram for 1980s. This shows continued progress in terms of increasing average life expectancies. The bottom 3 categories, from 20 to 50, had 49 countries, and now they only have 33 – most countries have life expectancies above 50. There an interesting exceptional data point: the category [20,30) was empty in 1970 – all countries had life expectancies above 30. But now there is one country in this category. Which country is it, and how did it fall back? Looking at the data set, we can identify the country as Cambodia, which went from a life expectancy of 42 in 1970 to only 29 in 1980. Only those who know the history of the Vietnam War, and the rise and fall of the Khmer Rouge would be able to understand why this happened – no amount of playing with numbers could reveal the meaning of this dramatic change.



The histogram is a tool of conventional statistics. It strips all data points of identity, and treats them all in the same way. The conventional methodology assumes that all data points are random draws from a common distribution. This makes them all homogenous, and makes it possible to treat them collectively. This is both a strength and a weakness of the conventional approach. It is certainly true that examining the data collectively can reveal interesting patterns. But the data also has a particularity to it – each point of data has an individuality and a personality – like the data of Cambodia. In “real” statistics, we pay attention to both the collective picture which homogenizes the data, and the particularities of each individual data point separately.

One way to do this, in the context of the present data set, is to look at the top ten countries and how they have changed across time.

| 1960s Top Ten | 1970s Top Ten | 1980s Top Ten | | | |
|----------------|---------------|----------------|------|-------------|------|
| Norway | 73.5 | Sweden | 74.6 | Iceland | 76.8 |
| Iceland | 73.4 | Norway | 74.1 | Japan | 76.1 |
| Netherlands | 73.4 | Iceland | 73.9 | Netherlands | 75.7 |
| Sweden | 73 | Netherlands | 73.6 | Sweden | 75.7 |
| Denmark | 72.2 | Denmark | 73.3 | Norway | 75.7 |
| Switzerland | 71.3 | Switzerland | 73 | Switzerland | 75.5 |
| New Zealand | 71.2 | Canada | 72.7 | Spain | 75.3 |
| Canada | 71.1 | Cyprus | 72.6 | Canada | 75.1 |
| United Kingdom | 71.1 | Spain | 72 | Cyprus | 74.8 |
| Australia | 70.8 | United Kingdom | 72 | Hong Kong | 74.7 |

An interesting pattern not in the numbers is the Nordic countries – Norway, Netherlands, Switzerland, Sweden, Denmark, Iceland in Northern Europe are highly over-represented. What are the similarities among these countries which lead to high life expectancies? Many potential factors are worth investigating. The appearance of Cyprus in the list is of interest, because it is

not a high income country. What is special about Cyprus that allowed it to enter the ranks of the top 10? What does it do differently from other countries in similar circumstances?

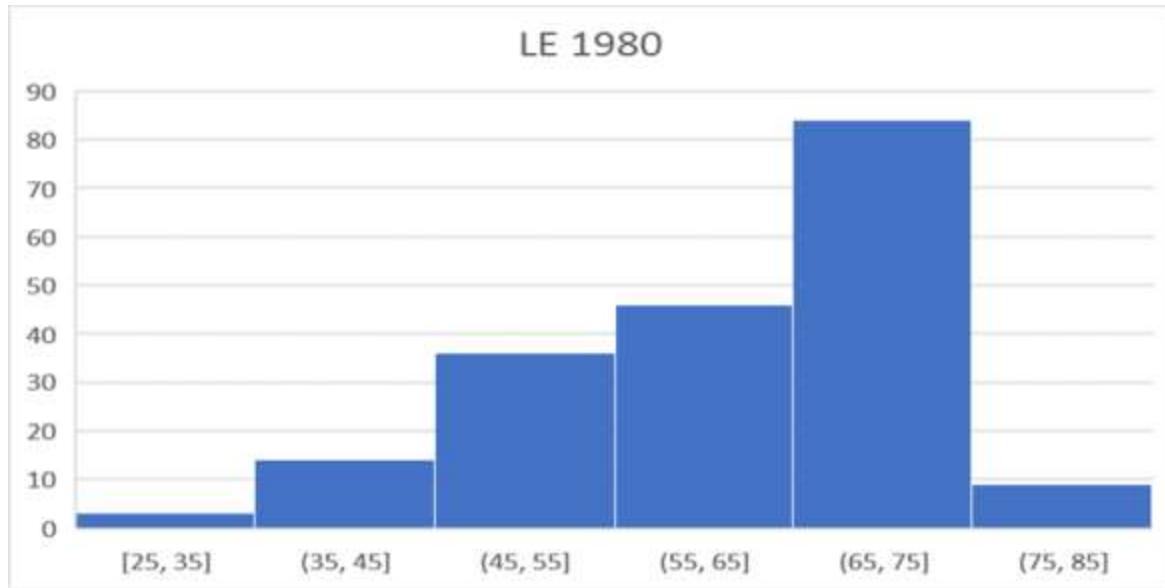
Above table shows the life expectancies of the top ten countries for the 60s, 70s, and 80s. Next we look at the more recent decades.

| 1990s Top Ten | 2000s Top Ten | 2015 Top Ten |
|------------------|------------------|----------------------|
| Japan 78.8 | Iceland 81.1 | Greece 84.3 |
| Iceland 78 | Greece 80.9 | Iceland 83.8 |
| Sweden 77.5 | Israel 80.4 | Israel 83.7 |
| Canada 77.4 | France 79.8 | Macao SAR 82.9 |
| Hong Kong 77.4 | Macao SAR 79.7 | Switzerland 82.8 |
| Macao SAR 77.3 | Japan 79.7 | Montenegro 82.7 |
| Switzerland 77.2 | Canada 79.6 | France 82.5 |
| Australia 77 | Netherlands 79.2 | Japan 82.5 |
| Italy 77 | Australia 79.1 | Channel Islands 82.4 |
| Greece 76.9 | Malta 79.1 | Netherlands 82.4 |

The rise of Greece to top position is very surprising, especially because Greece has been suffering from severe economic difficulties since the Global Financial Crisis. We have to look for an explanation. In general, when you have surprising statistics, there are two possible sources for the surprise. One possibility is mistakes in measurement – we must look carefully at how these statistics are compiled, and look for sources of large errors. The second possible source is that there was really a change in the health systems in Greece which led to a jump from 76.9 in 1990 to 84.3 in 2015. One would have to look at the real world to find the answers. These aspects of real statistics – where the statistics raise questions about particular aspects of the real world, and where each data point has a unique personality – differ radically from conventional statistical methodology which treats all data as a random sample from a common distribution.

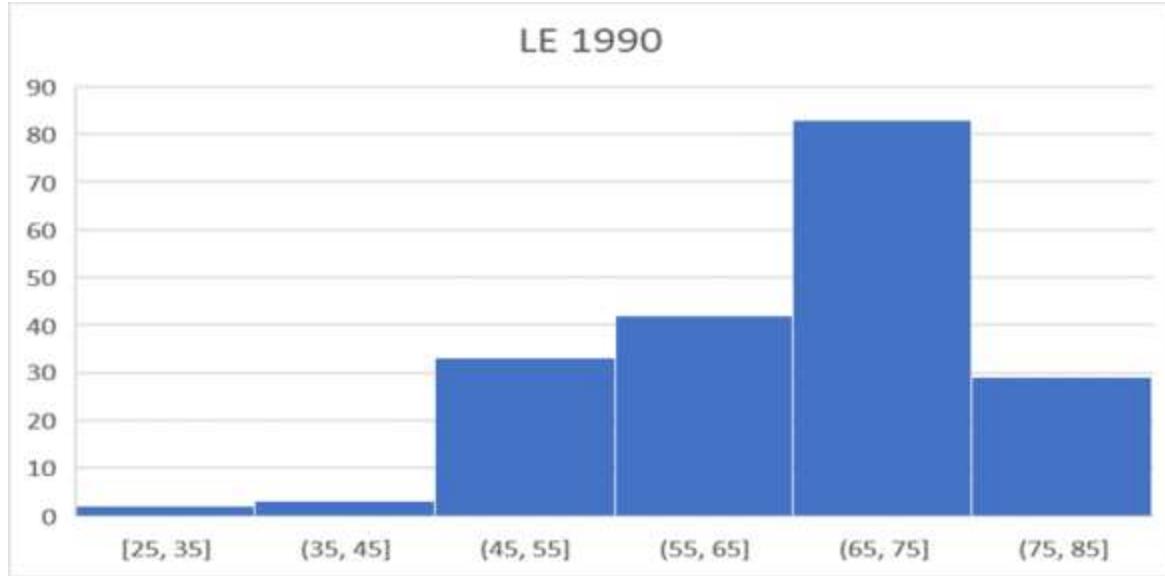
Analyzing the changes in the histogram over 1980, 1990, 2015

The next graph provides a histogram for the 1980s. We now see that countries have life expectancies in the 75-85 average range. Furthermore, the category of 65-75 is by far the biggest category. This is the modal category, with more than 80 countries in it. For the first time, two countries have gone above the 75 average and climbed into the 75-85 category, which was previously empty.



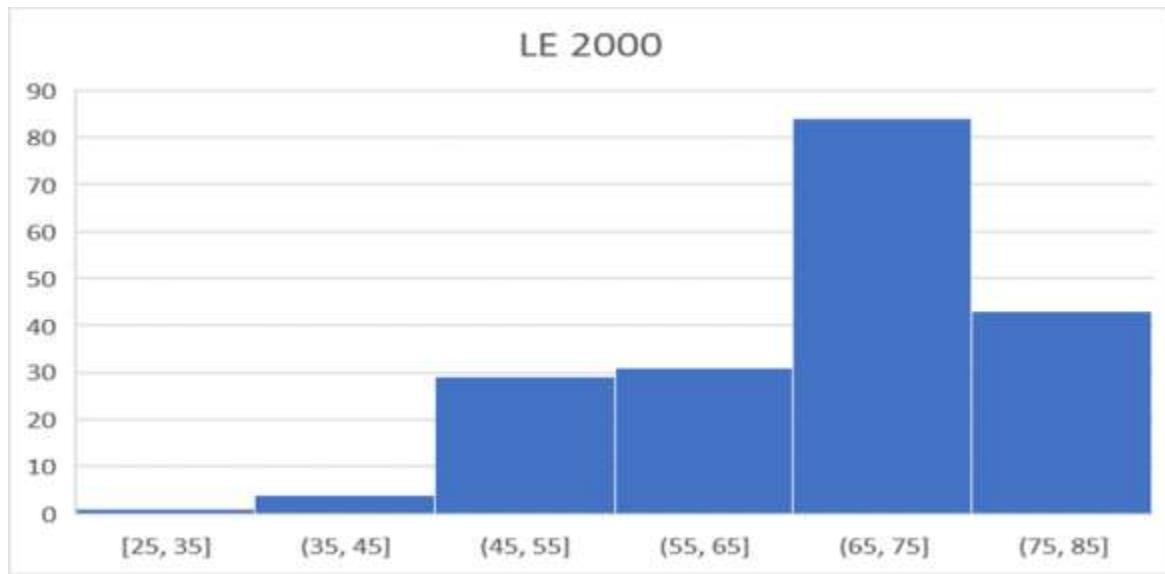
Computer Lab exercises associated with this chapter will teach you to extract the Life Expectancy for all countries in the data set in the 1980s, 1990s, and 2015. Note the changes in the top 10 and the bottom 10. Also, find the rankings of the particular countries we are following across time, and note any reversals in rankings – countries that were low in ranking moving up, and changing their positions. Think about WHY this could have happened, and note that the answer to this question will NOT be found in the data set. This is an essential aspect of “Real Statistics” – an alternative name for the Islamic Approach. The numbers are pointers to reality, and not the object of study. The observations provide us with clues about important aspects of reality, and further study in depth of the reality itself is required to follow up on these clues, and learn from the numbers.

Check your work on the Life Expectancy data set by comparing your histogram for 1990 with our histogram for 1990, which is displayed below



Similarly create a histogram for 1990. Sort data by Life Expectancy in 1990 to study the top 10 and bottom 10. Also note the relative rankings of your own country, as well as some other comparable countries – follow these rankings across time. Note the general rise Life Expectancy in the bottom countries? Why do we observe this phenomenon? Can this question be answered by any kind of data analysis?

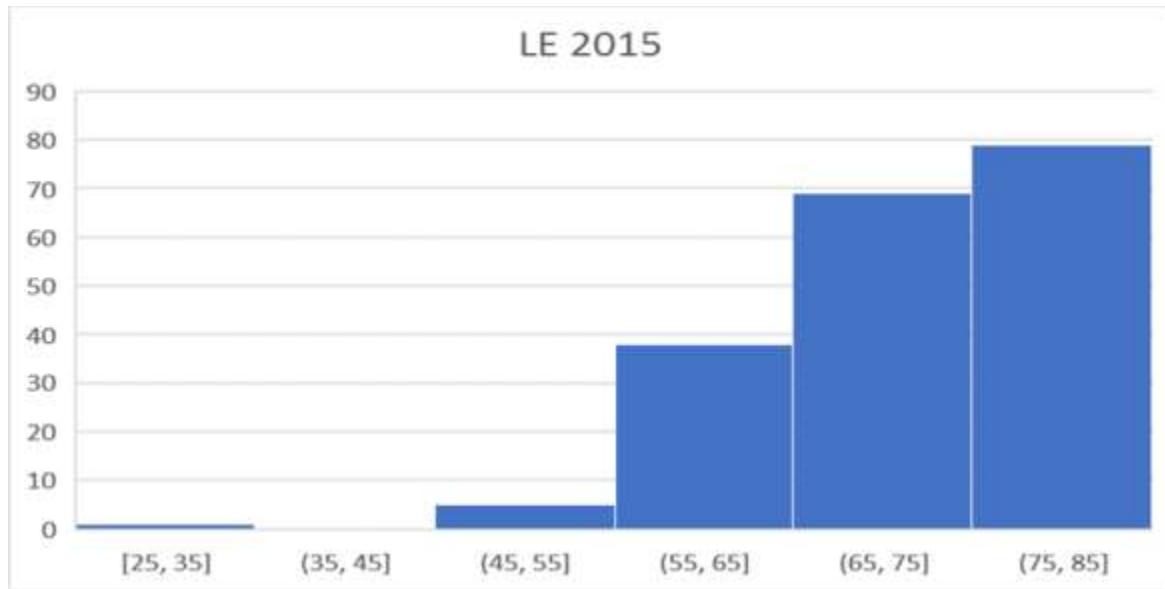
Next look at the histogram for the year 2000 plotted below:



Compute the rankings of the countries in 2000. This can be used to study the top 10, bottom 10, and the rankings of the particular countries we are following. There are many surprises in this data set. In particular, the strong performance of Greece is surprising. What accounts for the success of Greece in increasing life expectancies, despite serious economic

difficulties being faced by the country? To answer this, we need to have much more detailed information about Greece, and social policies which affected the mortalities in that country. Note that the increase in Life Expectancy must be due to a decrease in current mortalities in various age categories. How did the number of deaths go down, as a percentage of the population? Similarly note the behavior of the top 10 and bottom 10, as well as the other countries, and note any surprising reversals, as phenomena to explore further by studying the real world in greater depth.

Our last and final histogram is for Life Expectancies in 2015. For the first time, the category [75-85] is now the MODAL category, with the largest number of countries belonging to this category. Note that this category started out with ZERO countries in it in the 1960s. This shows that overall Life Expectancy has increased substantially over this period from 1960 to 2105, across the globe. The bottom two categories are now empty, and the lowest [45,55] has a tiny number of countries in it.



Concluding Remarks: The father of Western Statistics, Sir Roland Fisher, defined the subject as being the reduction of data. The key is to find “sufficient statistics” – a small number of ‘statistics’ which summarize the data. This is done by imposing theoretical assumptions on the data. For example, if we assume that the data is Normal, then the work of Fisher shows that two numbers – the mean and the standard deviation – carry the information contained in the entire data set. The point of summarizing is that we do not have the mental capacities to look at 10,000 numbers in the data set and understand them directly. However, today, with computers, an alternative approach is possible. Instead of imposing theoretical assumptions on the data, we can GRAPH the data in various ways to get a picture that we can understand. The object of Descriptive Statistics is to DIRECTLY visualize and understand the data. These pictures are to be used to get clues about the nature of the deeper reality which is manifested in the numbers. Life Expectancy is based on current mortality rates, and increases in LE correspond to reductions in mortality rates. To study WHY Life Expectancies have increased dramatically, we must study

the causes for the decline in mortality rates across the globe. These causes are not available in the data, and can only be discovered by qualitative analysis of features of the real world which have led to rising life expectancies.

3E Variable Bin Sizes

Preliminary Remarks: Mistaking the Map for the Territory

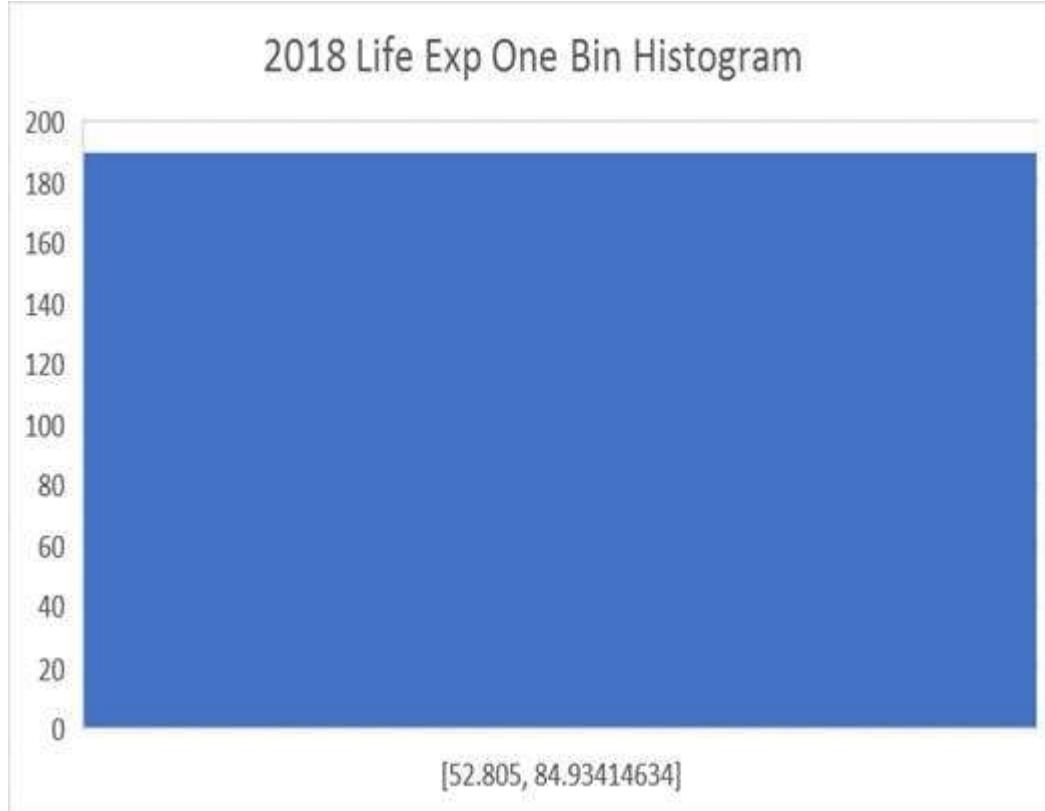
In order to be able to understand (simplify & reduce) data – it is useful to construct a statistical model for it. If the data follow a theoretical distribution, then they can be described by a formula. The DISTRIBUTION of the data may be identified with a theoretical distribution (like the Normal). If this is true, that allows us to substantially reduce the data set, since Normal distributions are completely characterized by only two numbers: the mean and the standard deviation.

A good way to identify the Data distribution is to look at the HISTOGRAM – a picture of the data. But, as we will see in this lecture, there are many possible histograms for the data, depending on the bin size. A traditional question is: What is the BEST model for the data? In the current context, what is the best bin size for making a histogram? This is the WRONG question. Data is primary, models are secondary. Different types of models describe different aspects of the data. As we decrease the bin size, we get a more refined picture of the data. At each level of refinement – histograms illuminate different aspects of the data. There is no one BEST bin size. We will illustrate this general concept by examining the histogram for Life Expectancies for 190 countries in the WDI data set for the year 2018.

We start by looking at the Default Histogram for 2018 Life Expectancy for 190 countries in WDI. The Histogram goes from MIN=52.8 to MAX=85.0 and makes 7 bins of equal size, where Bin Size = 4.6 years.

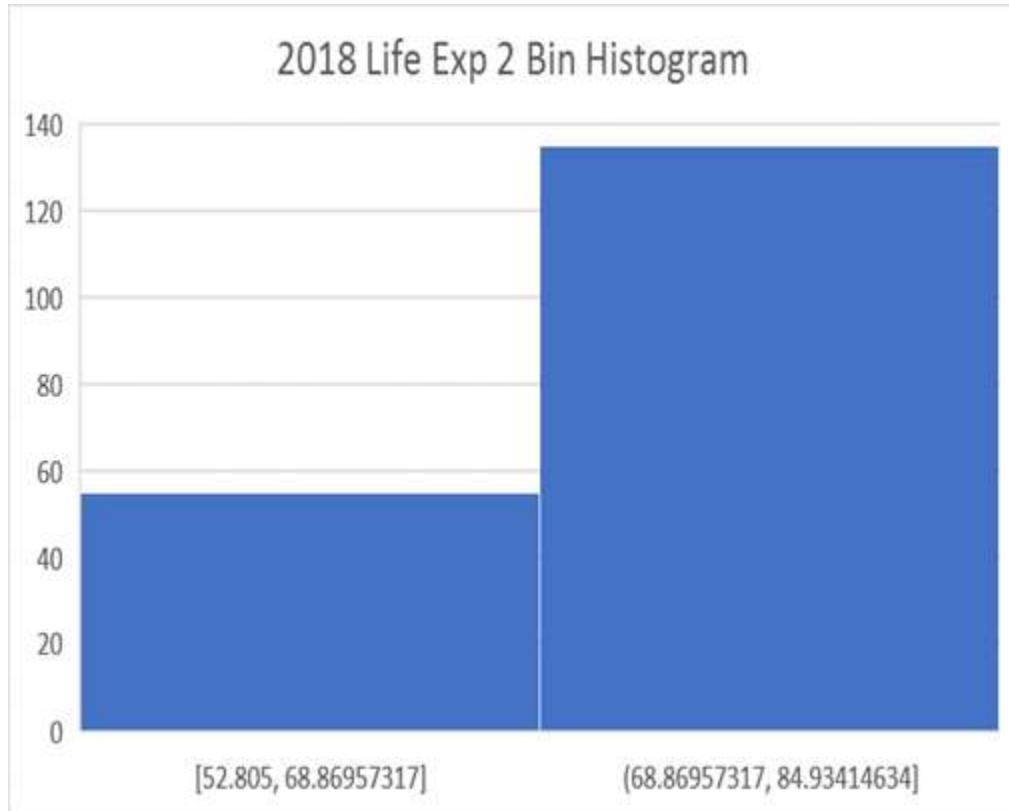
Starting with this as a baseline we will examine the effects of making the bin-size smaller or larger. In general, if bin size is too large, all data goes into one bin, and details are lost. On the other hand, if bin size is too small, every bin contains only one or zero data points and the groupings in data are not VISIBLE from the graph. The above 7 bins is a compromise between these two opposing effects, as we will soon see.

We start with the Coarsest Histogram with One Bin Only:



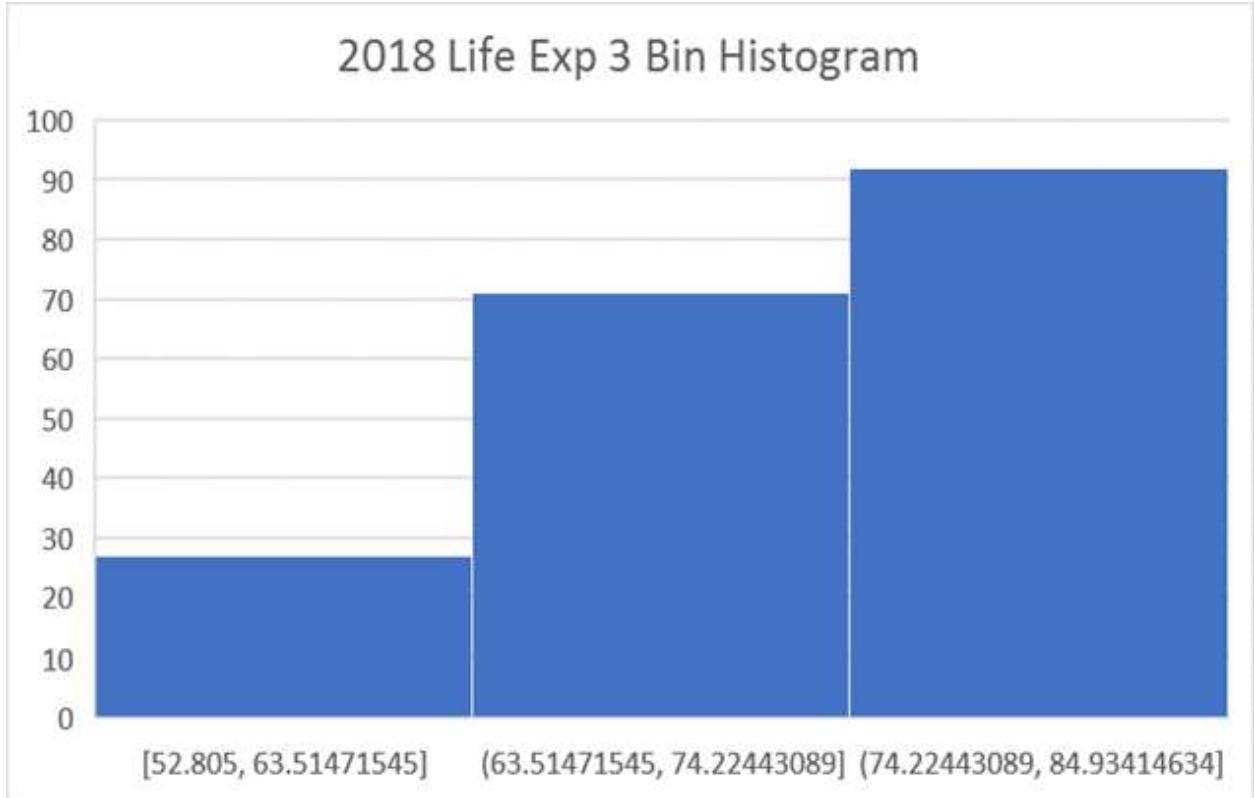
From this, we learn the RANGE of the data: it varies from MIN=52.805 to MAX=84.934. This is a COUNT histogram. We learn that there are 190 countries in the data set from the vertical axis. Later we will study a PERCENTAGE or PROBABILITY histogram. This gives us the proportion of the population in a given bin. From a probability histogram, we would not learn the count, since only 100% would appear on the vertical axis.

Next, let us look at a histogram with only two bins:

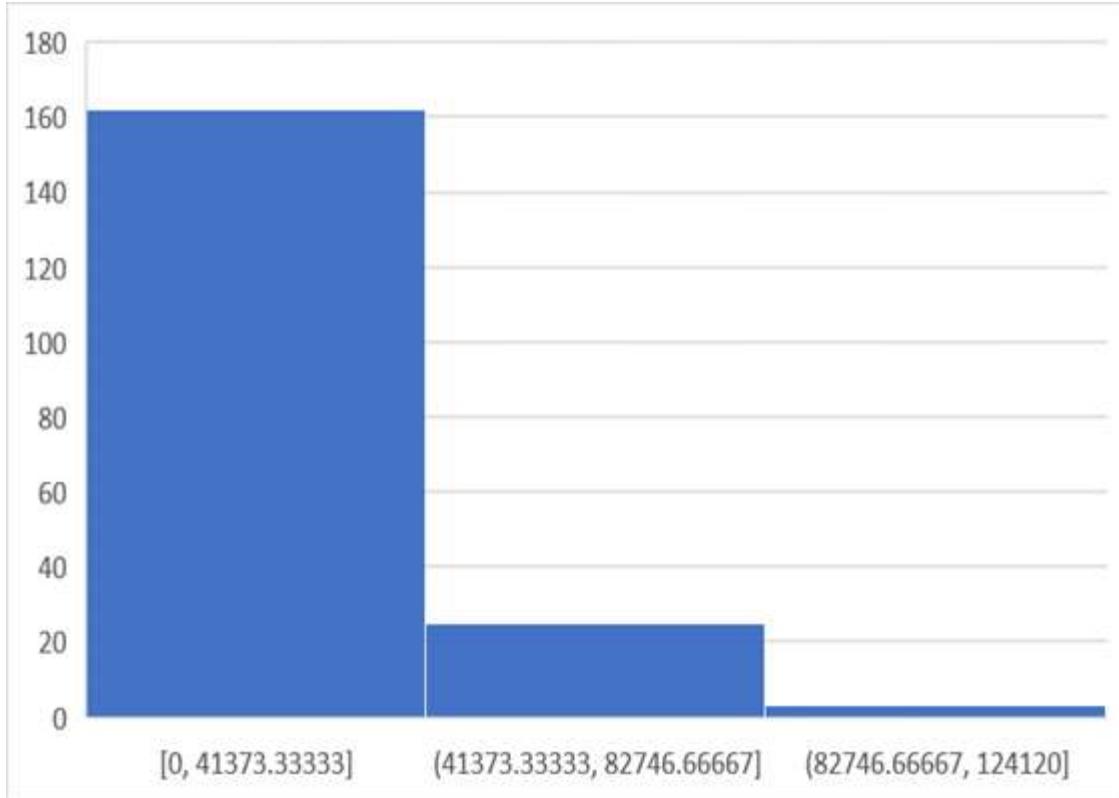


Two Bins divide the range from 52.8 to 84.9 into two equal parts. The midpoint of the range is 68.8. 55 countries are in the first bin of below midpoint Life Expectancy, while 135 countries in the 2nd bin. Clearly, the distribution is NOT symmetric. From this graph, it is obvious that the Normal distribution would NOT be the right model for this data set.

The 3 Bin Histogram divides countries into three categories – high, middle, and low Life Expectancy. The Low LE Bin goes from 52.8 to 63.5, and has only 27 countries. The middle LE bin goes from 63.5 to 74.2, and has 71 countries. The high LE bin goes from 74.2 to 84.9, and has 92 countries:

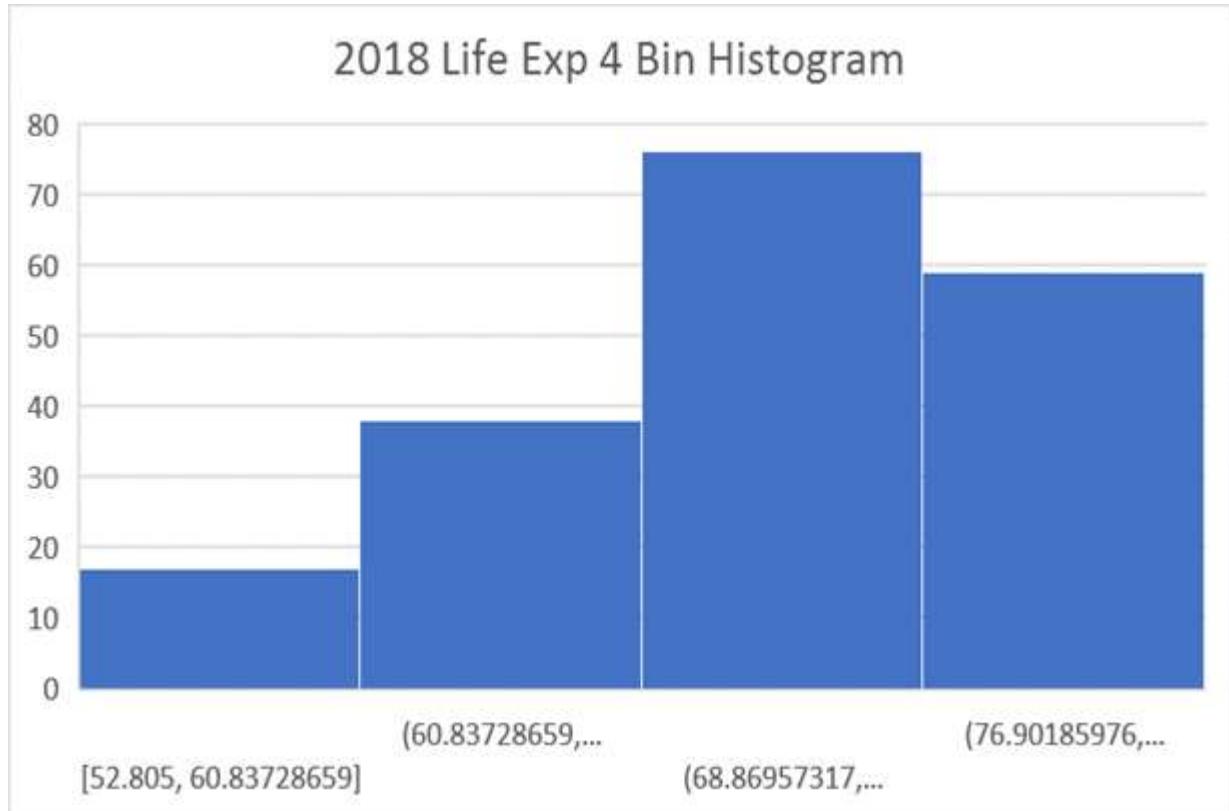


What is very surprising is that the largest number of countries are in the highest category. The MODE is the bin (or category) that has the largest number of categories. The graph shows that the Mode is at the last bin. WHY is this very surprising? That will become clear if we look at the histogram of these same 190 countries classified by GNP per capita in the same year 2018. This is graphed below



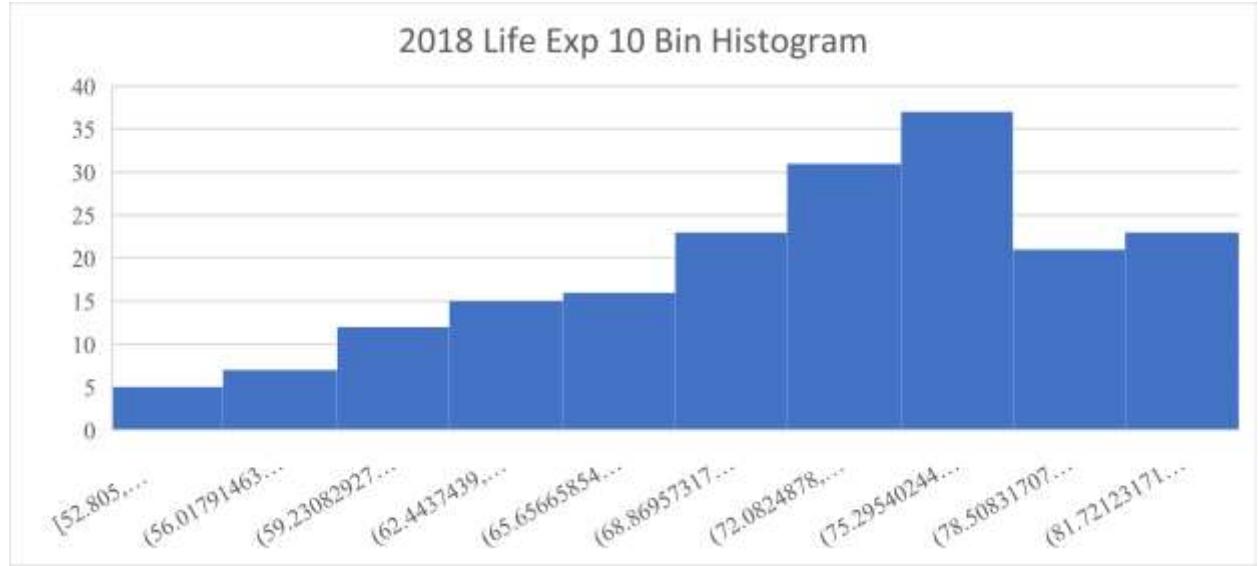
This 3 bin Histogram of GNP per capita, in PPP terms, constant USD, shows that only a few countries belong to the high GNP category, and the vast majority belong to the low GNP category. This shows that EVEN countries in the bottom third income category can achieve high life expectancies for the population. This means that cheap and simple measures are sufficient for substantially and significantly lowering mortality rates. One does not need to wait to grow rich as a country, in order to take effective measures to improve the health of the population.

The 4 Bin Histogram divides the range into four categories, with Bin Width = 8 years:

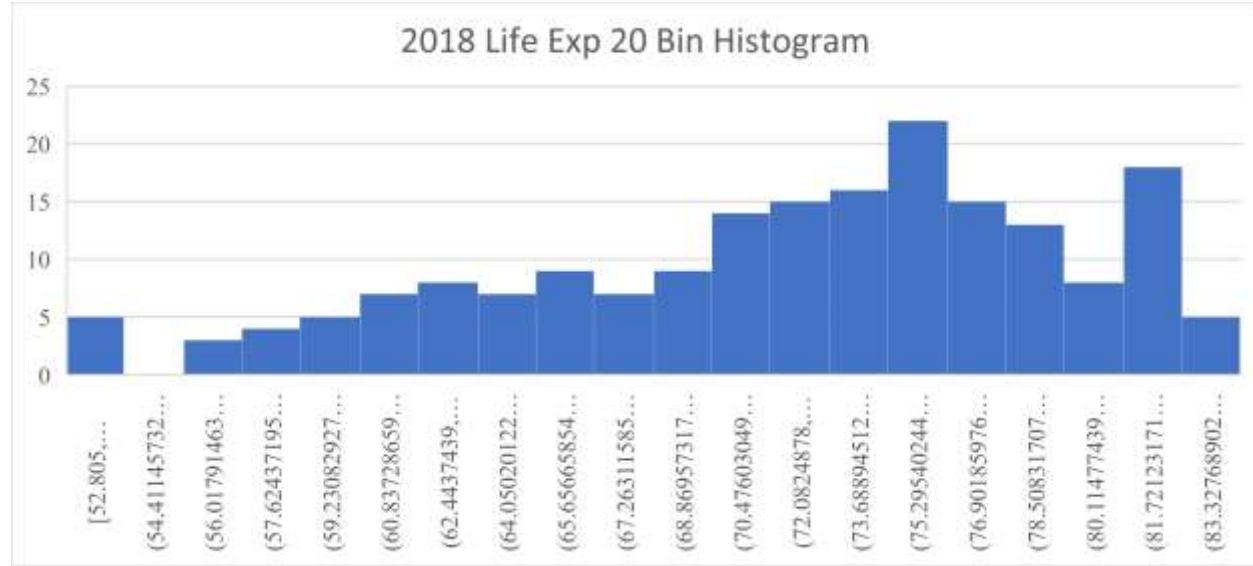


In this histogram, the Modal Bin is [68.8, 76.9] with 75 countries. There are only 59 countries in the highest bin, going from 76.9 to 84.9. The graph suggests that it is relatively easy to get LE up to 70, much harder to get it up to 80. To learn more about this, we need to look at the mortality rates in each age group. By comparing between countries with low and high mortality rates, we can learn about where is the greatest potential for improvement. To realize this potential, we need to investigate carefully the causal determinants of mortality.

As we go through graphs of 5, 6, 7, 8, and 10 bins, we get more information about how the data divides into different kinds of groupings. The 10 bin graph is plotted below:

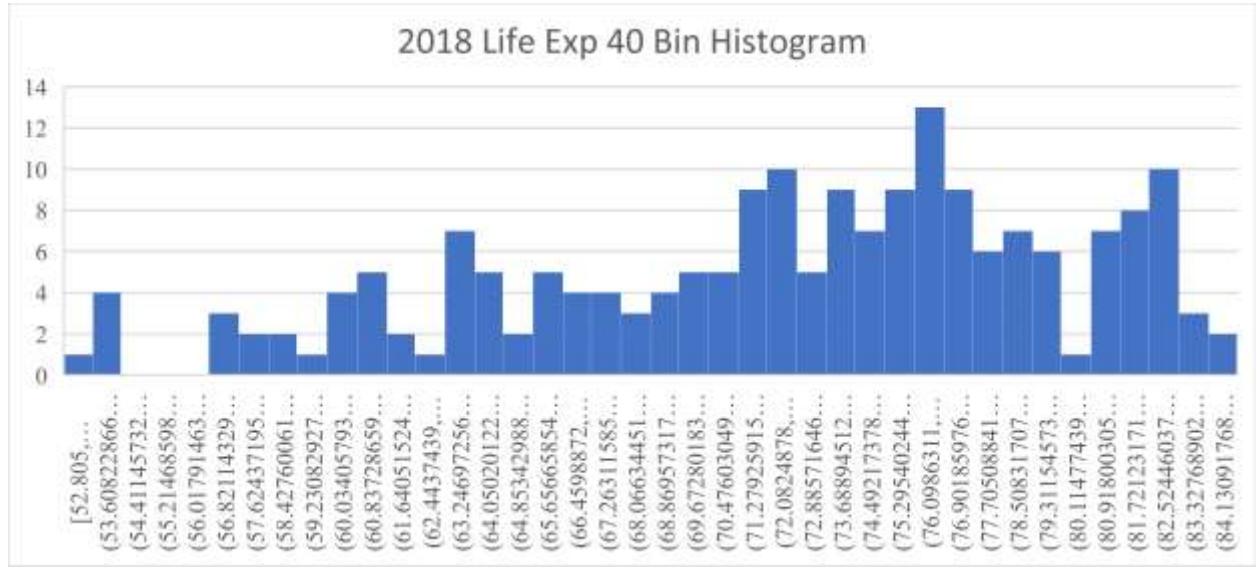


At each level of refinement we get more information about the data, and we also pick up some visual patterns not visible at other levels of refinement. However, as we increase the level of refinement, we start losing the ability to look at the graph and interpret it directly and visually. Here is the histogram with 20 bins:



There are **FOUR** modes in this histogram. When the number of countries in a bin is small, countries can fall into a bin or out of it by statistical accident. When you have two bins, High and Low, classifications would be robust to small errors – regardless of how you compute it, the classifications would remain the same for most countries. However, when you make up a large number of categories, this is no longer true, and classification can be much affected by small errors in the data. Thus the number of countries displayed in the graph is **NOISY** – it is

much affected by errors. As we make the bins even smaller, the noise increases even more and the patterns in the data are no longer visible. The graph with 40 bins is very jagged and noisy, and the broader patterns are barely visible:



becomes visible AFTER we make the graph, so choice of “optimal” bin size is impossible. The choice of bin size gives us a Distribution that provides a MODEL for data, but there is no TRUE model. All models are approximations to enable us to summarize the data and understand it.

A deeper understanding requires an examination of mortality rates and their causal determinants. This requires going further, beyond the data sets, into examining mortality rates, classifying them by type, and examining causes of each type. Numbers give us clues about the real world, but are never the goal of the analysis. Statistical analysis must be followed up by examining real-world issues that they highlight.

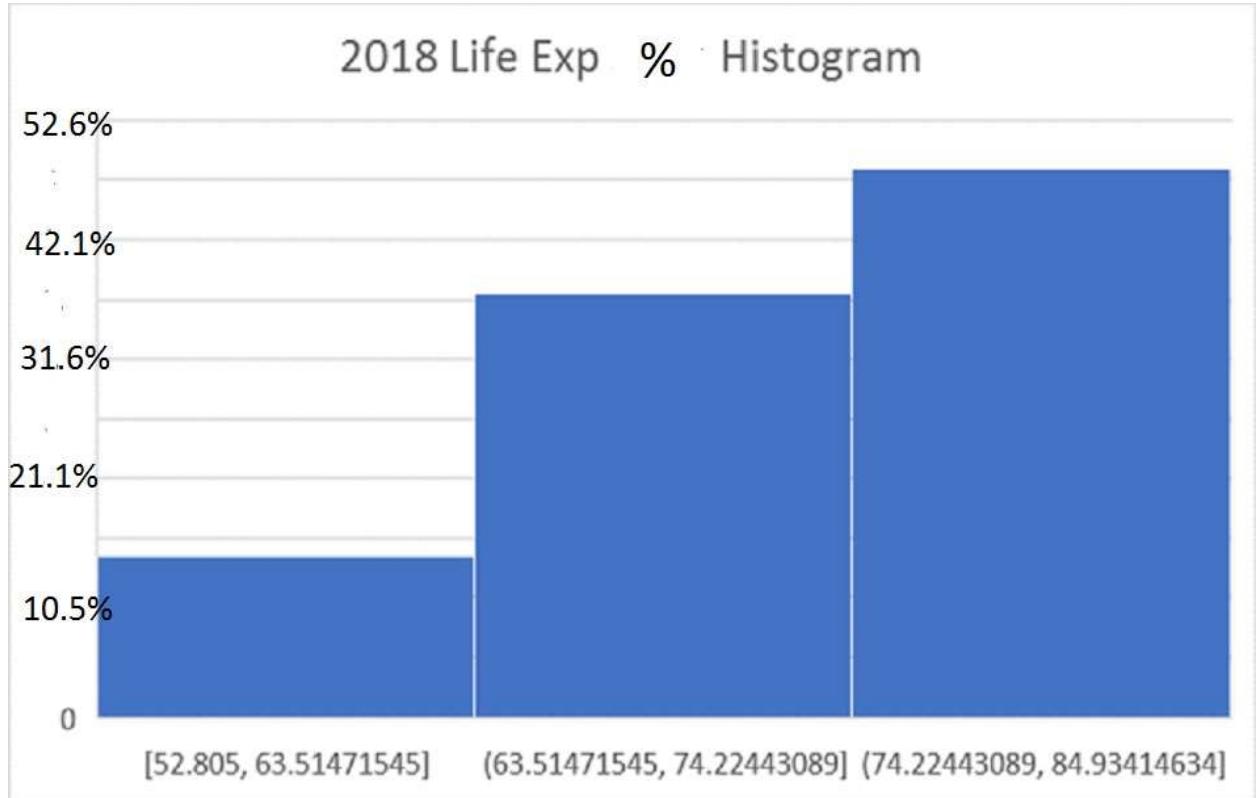
3F Probability Histograms and the CDF

The goal of this course is to teach about simple descriptive statistics, which allow us to look at and understand the data set. The central link between this, and more advanced concepts, is the random sample. Part E of Lecture 3, explains some of the most elementary concepts in connection with random samples from populations of objects in the real world.

Concept 1: Choosing one member of the population “at random”. A dictionary defines “random” as: “made, done, happening, or chosen without method or conscious decision”. For example, I could choose a country at random by throwing a dart at a map of the world. However, statisticians use the word “random” in a technical sense, which is very different from the standard English language usage. Choosing at random means that all countries in the population must have exactly equal chances of being chosen. Throwing a dart does not allow us to calculate the probabilities of selection for each country because there are too many variable and subjective factors, including the fact that different countries have different sizes. A reasonable way to do random choice among the 190 countries in the WDI list is to choose a random number between 1 and 190 – for example, via the EXCEL function RANDBETWEEN(1,190). This function gives an equal chance to all numbers and if repeated often, will make all numbers turn up equally often in large samples. It is common to confuse the two senses of the word, the haphazard choice of English language (**E-random**), and the systematic choice which give equal probability to all possibilities of Statistics (**S-random**).

Concept 2: To relate S-random choice to Histograms, consider a three-bin histogram for the 190 countries, discussed in the previous lecture. The LOW bin goes from 52.8 to 63.5, and has 27 countries. The MID bin goes from 63.5 to 74.2 and has 71 countries. The HIGH bin goes from 74.2 to 84.9, and has 92 countries. The probability Histogram answers the following question: If a country is chosen at random from this population of 190 countries, what is the PROBABILITY that it belongs to any one of the three bins? The answer is obvious. The LOW bin has 27 countries, and so a probability 27/190 of being chosen. The MID and HIGH bin have probabilities 71/190, and 92/190 respectively. This histogram is pictured below. It is EXACTLY the same as the 3-Bin histogram in the previous lecture (rsra03E) EXCEPT for the labels on the Vertical Axis. The COUNT histogram gives us the count of the number of countries within each category. The PROBABILITY histogram replaces the count by the Percentage of countries within each bin. In the COUNT histogram, the axis labels were 10,20,...,100, corresponding to the number of countries. In the current Probability histogram, these numbers have been replaced

by percentages corresponding to $10/190=5.3\%$, $20/190=10.5\%$, $30/190=15.8\%$, ..., $80/190=42.1\%$, $90/190=47.4\%$, $100/190=52.6\%$. The number of countries has been divided by 190, the total number, to give percentages in each category,



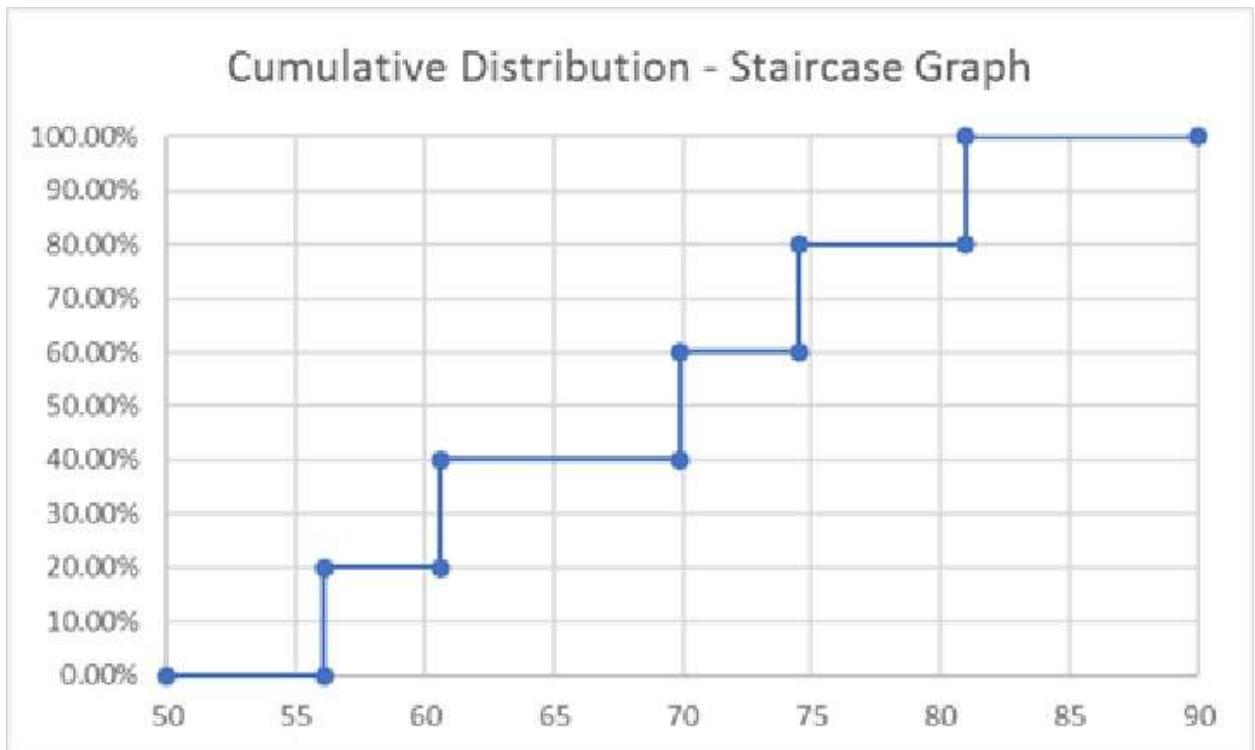
Let \mathbf{C} be a randomly chosen country from among the 190 countries in the sample. Then the probability histogram above plots the following three probabilities: $\mathbf{P}(52.8 \leq \text{LE}(\mathbf{C}) \leq 63.5)$, $\mathbf{P}(63.5 < \text{LE}(\mathbf{C}) \leq 74.2)$, and $\mathbf{P}(74.2 < \text{LE}(\mathbf{C}) \leq 84.9)$. These are the probabilities of the randomly chosen country having life expectancy within the range which defines the bins. These probabilities show the “distribution” of the random variable Life Expectancy – the height of the graph tells us the chance of the random variable being in the specified range. For theoretical purposes, the CUMULATIVE Distribution turns out to be a very important concept. Instead of looking the probability of a bin – $\mathbf{P}(a < \text{LE}(\mathbf{C}) < b)$, the cumulative distribution looks at ALL the probability upto a given number: $\text{CDF}(x) = \mathbf{P}(\text{LE}(\mathbf{C}) \leq X)$.

To explain the cumulative distribution function, it is useful to look at a case where there are only a small number of countries, as it makes the concepts easier to understand and visualize.

| | | | |
|-----|--------|-----|------|
| 25 | Qatar | QAT | 81.0 |
| 84 | Turkey | TUR | 74.5 |
| 125 | Egypt | EGY | 69.9 |

| | | | |
|-----|-------------|-----|------|
| 169 | Afghanistan | AFG | 60.6 |
| 183 | Sudan | SDN | 56.1 |

Assume that these five countries, chosen “E-randomly” (in the English language sense) from the 190 countries, are the full population of countries. We will choose a country at S-random from these 5, and call it **C**. $LE(C)$ represents the Life Expectancy of the randomly chosen country. A characteristic of a randomly chosen country, like LE , is called a random variable. For any real number x , we want to measure the $P(LE(C) \leq x)$ – this is, by definition, the cumulative distribution function of the random variable $LE(C)$. This can easily be plotted as follows.



For all values of x less than 56.1, the probability that the Life Expectancy of the randomly chosen country is less than this x is 0%. The graph shows that $P(LE(C) \leq x) = 0\%$ in the first step of the graph. This is because Sudan has an LE of 56.1, which is the smallest LE in the data set. For values of x between 56.1 and 60.6, $P(LE(C) \leq x) = 20\%$. There is only one country (Sudan) that has LE below this range of values. The chances of randomly choosing Sudan are 1 out of 5 or 20%. Similarly, for values of x between 60.6 and 69.9, $P(LE(C) \leq x) = 40\%$. For values of x in this range, there are two countries, Sudan and Afghanistan, which have LE 's below x . The chances of choosing either one of these two countries from a random choice between 5 countries are exactly 2/5 or 40%. Similarly, the graph jumps by 20% at the Life Expectancy of each of the five countries in the data set. This brief description provides a preliminary introduction to the concept of the cumulative distribution function. More details will be discussed in a subsequent course, where the ideas of random sampling, which are at the heart of statistical theory, will be developed. In the current course, we are only concerned with

different graphs which represent the data in different ways. The CDF of the data, pictured above, is also one of these ways to make a graph of the data.

Chapter 4: Reducing Data to One Number

Begin a new chapter here...

4A: Inflation: A One Number Summary of Changing Prices

This topic is generally discussed under the heading of “Measures of Central Tendency”. The idea is to represent the entire data set by just one number. Typically this number will be a ‘central’ number within the data set. There are many different ways to formalize the concept of “central”. We reject this approach, because this type of data reduction cannot be done without knowing the purpose of this reduction. To learn the purpose, we need to go beyond the data set, to the real world problem we are trying to solve, with the help of the data set. We will illustrate this issue by discussing “inflation”, which is one number that represents a lot of different price changes.

The Traditional Approach to data reduction was created by Sir Ronald Aylmer Fisher, known as the father of modern statistics. He assumes that the data are generated from “nice” theoretical density, like the normal density. If this is true, then the best use of the data is to discover the density. It is possible to show that the Data = Theoretical Density Plus Random Errors (Noise). We can reduce data by eliminating noise. Under convenient ASSUMPTIONS about underlying density which generates the data, Fisher showed that we can find SUFFICIENT STATISTICS – one (or more) numbers which provide ALL the information that data has about the underlying density. This provides a theoretical justification for taking a large data set and reducing it to just a few numbers.

However, WHAT if data is NOT generated by underlying density? Then it cannot be reduced to one number, and we must deal with the entire data set. What is the JUSTIFICATION for making the assumptions which allow us to reduce the data to sufficient statistics? The ONLY justification is that it allows a theoretically valid reduction. We can NEVER prove that the assumptions regarding nice underlying density are valid. In the pre-computer era, it was convenient to make assumptions which allowed reduction of the data, but there is no longer any need to make fairy-tale assumptions just for convenience, so that we do not have to deal with large amounts of data.

What is the ALTERNATIVE to making assumptions? This involves Narratives – using data to tell coherent, explanatory, causal, stories. We will discuss how this process works in this course.

An Islamic Approach involves looking through the appearances to recognize the underlying reality. In particular, we want go beyond the numbers to the real world which is being measured by these numbers. In the context of data reduction, the questions of importance are:

1. WHY do we want one number to describe the entire data set?
 2. Assuming this is needed, HOW can we compute this number?
 3. WHAT does this number mean – with reference to data, AND with reference to the real world?
- TO elaborate the question 3 above, REAL WORLD generates DATA which is reduced to One Number. What is the relationship of the one number to the real world?
4. What will the effect of this data reduction, replacing the whole data set with one number?

Before explaining the answers to these questions in the context of inflation, we explain a general Pedagogical Principle: Understand Abstraction by reduction to CONCRETE. When we have Theorems, Principles, Philosophies, we can only understand them by asking: HOW do these apply in simple real world examples? It is necessary to take abstractions and understand how they work in the real world, in particular special and SIMPLE examples. By understanding SEVERAL such applications, one gets an idea about how the general principle works. In this lecture, we study INFLATION – ONE number to represent hundreds of changes in prices. What does it MEAN to talk about the general level of inflation, when prices of different things changes in different ways. To understand this general question, we look at a specific example on a real data set. There are twelve general categories of goods which are used to calculate the inflation numbers. These are listed below.

| Category | May-19 | May-20 | |
|--|--------|--------|-------|
| Food and non-alcoholic Beverages | 118.58 | 131.55 | 10.9% |
| Alcoholic Beverages, Tobacco | 121.92 | 145.03 | 19.0% |
| Clothing and footwear | 120.26 | 131.67 | 9.5% |
| Housing, water, Elec., Gas and other fuels | 122.23 | 130.22 | 6.5% |
| Furnishing and household equipment maintenance | 119.56 | 128.8 | 7.7% |
| Health | 128.7 | 139.92 | 8.7% |
| Transport | 133.61 | 124.89 | -6.5% |

| | | | |
|---|--------|--------|-------|
| Communication | 105.92 | 108.22 | 2.2% |
| Recreation and culture including Stationery | 112.91 | 118.12 | 4.6% |
| Education | 142.32 | 144.98 | 1.9% |
| Restaurants and hotels | 121.07 | 129.8 | 7.2% |
| Misc. goods and services | 123.55 | 137.02 | 10.9% |

A basket of goods representing average household expenditure on that category is priced in May 2019. For example, the chosen food basket costs PKR 118.58 in 2019. The same basket is priced in May 2020 and now costs PKR 131.55. This is an increase of 10.9%, which is the inflation rate for the food category. Note that by changing the composition of the food basket we could change this rate. The same kind of calculation is done for each category of goods, so that we get 12 different rates of inflation, one for each category. These rates are listed in the last column. The highest rate is 19% for Alcoholic Beverages and Tobacco, while the lowest rate is -6.3% for Transportation. This negative rate is due to the unusual behavior of oil prices, which decline over this period of time. To get a SINGLE number, we could just average these 12 numbers. If we do this, we get 6.9%. However, it does not seem reasonable to take a simple average, because some of these categories are far more important than others. Using weights we will discuss in much greater detail later, we find the Rate of Inflation is 7.3% Year-on-Year (YoY) basis from May 2019 to May 2020. We could also look at MoM Month-on-Month inflation by comparing price changes on a monthly basis. There are many different ways to come up with an inflation number. Before we can proceed further, we must ask: What does this number (7.3%) MEAN? In order to understand the meaning, we must look at the following questions:

1. How is this 7.3% computed from this table? As we will see, this is based on a choice of weights, which is somewhat arbitrary, but not entirely so.
2. Is it REASONABLE to even TRY to reduce such a diverse collection of price changes to ONE Number? What is the compulsion to do such a reduction? Why cannot we just use 12 different rates of inflation, one for each group?
3. What is the EFFECT of making such a reduction?

As we will see, even though there are many arbitrary decisions that go into the calculation, this number (7.3%) is called the HEADLINE Inflation! This number enters into many different decisions of economic policy. Before crunching numbers, we need to know WHY? WHY do we want ONE number to represent/summarize all of these movements? What are the EFFECTS of reducing complexity and diversity, and condensing all price variation into a single number: the headline inflation?

Answers to WHY questions have many dimensions. One dimension is the history: why did people start calculating these numbers? Why did people think we could find one number which would represent the general price level and the general rate of increase in prices? To answer this, we need to discuss some Economic Theory. The Quantity Theory of Money (**QTM**): Suppose we add a ZERO to all notes. The quantity of Money is now multiplied by 10. ALSO, all prices are multiplied by 10. But, there is no change in the real economy.

QTM: Suppose MONEY STOCK grows by 15% and there are no REAL changes in the economy. THEN all prices will increase by 15%, so the rate of inflation will be 15%.

If all prices rise roughly in the same proportion, then inflation makes sense. If inflation is occurring due to increasing quantities of money, then we would expect to see this pattern. However, our data set shows that this is NOT the case. Even though the rate of inflation is calculated to be 7.3%, inflation in different categories of goods ranges from a high of 19% to a low of -6%. So the question arises:

Can we use ONE number (7.3%) to REPRESENT, or STAND for, many numbers – the whole set of inflation numbers from 19% to -6% across the different categories of goods? Again, this depends on the PURPOSE for which we are doing the representation.

Inflation numbers serve many purposes, among which THREE BROAD Purposes can be identified as follows:

1. Evaluating Government Performance (also SBP, Treasury, Monetary Policy). Governments attempt to provide stability of prices, and keeping inflation low is a policy target.
2. Making Monetary Policy DECISIONS: The inflation number is very important in deciding upon appropriate monetary policy. The State Bank of Pakistan (and all Central Banks) look at a broad range of inflation forecasts, and use them in MODELS of monetary policy
3. Public plans for the future depend on inflation forecasts. Consumers worry about inflation because they need to use money to purchase goods. The business Sector needs to adjust the wages of laborers to keep in line with inflation, and to set prices so that they cover rising costs due to inflation. The financial sector must set interest rates to cover inflationary costs. For all of these purposes, inflation numbers are needed. But can one number be suitable for all, or would different numbers be useful for different purposes?

In fact, many different inflation numbers are computed for different purposes. There are Consumer Price Indices, Wholesale Prices for Business, Core Inflation numbers, and many other types. In the remaining lecture, we will discuss ONE of these numbers in greater detail: the Sensitive Price Index or SPI. What is the purpose of the SPI? Actually, there are many:

1. Judge impact of inflation on the general public. In Pakistan, the masses are poor and need to buy essential goods, so inflation in these goods affects their lives more.

2. Big changes in SPI lead to Political Unrest, loss of Political Support for the party in power, and have an impact on Election outcomes.
3. Labor demand for wages, and salary increases of Government Servants require consideration of inflation in essential goods.
4. Many types of Social Service programs provide money to the poor. The question of “How much money should be provided” requires knowing the prices of essential goods.
5. If we want to know the level of income required to be able to purchase basic needs, we need to know about the SPI.
6. Where to draw the poverty line, such that those below the line are counted as poor and eligible for support? There are many possible uses and purposes for a poverty line, and these also impact on how we should compute the SPI numbers.

How can we adjust the SPI to take into account these different concerns? There are two main areas where we have some flexibility. We can make a CHOICE of goods, and also a CHOICE of weights. We will show how the actual calculations are done in the next lecture. We end this lecture with some **concluding remarks**:

1. There are many different commodities, with many different prices. Prices increase at different rates in different sectors.
2. “Inflation” measures – using ONE number only – ALL of these different changes.
3. This cannot be done – one number cannot represent all of the different price changes.
4. If we want to use only one number to measure inflation, we must know the PURPOSE for which this data reduction is being done. Given a specific purpose, some suitable strategy for data reduction may be possible.
5. Standard statistical theory makes arbitrary theoretical assumptions, impossible to verify, in order to reduce data. This does not seem like a useful method for real-world purposes. It was created in the pre-computer era because it was impossible to analyze large data sets without making such simplifying assumptions.

4B: The Sensitive Price Index

Before turning to details of how a price index is computed, we discuss some general principles involved in the acquisition of knowledge. There is an important distinction between the Analytic Versus Synthetic approach.

1. **Analytic Approach:** Break everything into small parts, and study them separately. Part of a divide and conquer strategy towards knowledge.
2. **Synthetic approach:** put all the separate pieces together to study them as a whole. Part of a holistic and integrated strategy for achieving global understanding.

In general, we need both approaches for acquiring a good understanding. We need to understand the pieces separately, and also how they work together. In general, the Western intellectual tradition is very strong on the analytical approach, and very weak on the synthetic. The most important loss from this weakness occurs in the context of EMERGENT properties. These are properties of the system as a whole that cannot be understood from studying the parts of the system in isolation. For example, studying the properties of heart cells in isolation cannot lead to an understanding of HOW, and much less WHY, the heart pumps blood. The WHOLE is greater than the sum of its parts. The standard Approach to Statistics is isolationist and analytical. Islamic approach is HOLISTIC and synthetic, and starts with consideration of PURPOSE.

A price index reduces a large number of prices to ONE price. This is one type of data reduction. All Data Reduction HIGHLIGHTS something and IGNORES many things. To understand data reduction, we must understand “What is being highlighted?” and also “What is being ignored?”. We must understand the PURPOSE of data reduction, in order to be able to highlight the right elements, and to understand what can safely be ignored.

The Sensitive Price Index (SPI) is meant to study how expensive it is for the masses to buy necessities. Increases in this index reflect increased difficulty in buying essentials. This general understanding is of great importance in deciding upon many details which come up in constructing the index. CLEARLY, one should pay attention to public provision of necessities – like social welfare programs, government hospitals, government provision of education. ALSO, we should pay attention to the availability of non-market goods, like self-grown food (in rural areas). All of these factors matter in terms of how much money is needed by the masses to enable them to buy essentials. However, these factors would not be considered in a standard statistics course. Going outside the study of numbers is strongly discouraged by the conventional approach, which is based on an analytical approach. The statistician's job is confined to the analysis of numbers, and not to the bigger real-world context from which the numbers come.

After these preliminary remarks, we turn to the technical details of the calculation of a price index. The first step involves choosing a Bundle of Goods, which are “representative”. In real life, I was part of a committee of experts at the Pakistan Bureau of Statistics (PBS), which was called to decide on goods to be put in the new commodities basket for 2015. We had an initial list prepared by the PBS and subjectively modified it by adding and subtracting commodities the member felt were important. The larger the collection of commodities, the more representative they would be. But also, the process of gathering information on prices, and also on expenditure, would be more costly and time-consuming. So it was important to pick a small bundle, but one which would also be representative.

Turning to the specific details of the SPI, we note that there are 53 goods in the basket for the SPI. These goods are listed below:

Looking at the list shows that goods reflect commodities of importance to the poor in Pakistan. But it does not, and cannot, reflect regional and seasonal variations, and non-market effects. For example, firewood is important for heating in the lives of the poor in cold regions,

but is not sold in normal marketplaces. People in rural areas can grow some of their own food. Most importantly, rent/housing is not included, even though it is a major component of the cost of living. This is because of CONVENTIONS regarding the Consumer Price Index in Pakistan. In other places, rent is included. For example, COLA is short for Cost of Living Adjustments made to retirement incomes in the USA, and this includes rental prices.

Coming back to the details of calculations. The first step, choosing representative commodities for the PURPOSE of the index, has already been discussed. The second step is to Get the Prices of Chosen Commodities. We illustrate this by looking at a subset of 11 commodities chosen from among the 53 in the SPI list:

| | | 25/10/18 | |
|-------------------------------|---------|----------|--|
| Item | QTY | Price | |
| Wheat Flour, Bag | 10 kg | 182.8 | |
| Rice Basmati Broken, (AQ) | 1 kg | 37.8 | |
| Chicken Farm, Broiler, Live | 1 kg | 83.4 | |
| Milk, Fresh, Unboiled | 1 Ltr | 30.5 | |
| Cooking Oil, Tin, (SN) | 2.5 Ltr | 316.3 | |
| Pulse Masoor, Washed | 1 kg | 71.4 | |
| Potatoes | 1 kg | 15.2 | |
| Sugar, Refined | 1 kg | 27.9 | |
| Tea Prepared, Average Hotel | Cup | 6.9 | |
| Electricity Charges (Average) | Unit | 4.4 | |

Why did we choose 10Kg for Wheat, and 1Kg for Rice, 1 cup of Tea? The UNITS chosen are ARBITRARY. We will price the SAME bundle across time and look at changes. It is only the percentage change in prices that matters. So we choose UNITS on the basis of what is convenient and easy to price in the market. In general, there are many prices for the same good, so we need to make conventions about WHAT counts as THE price, and how to handle fluctuations in prices. It does not matter how we do this, as long as we choose a systematic method that remains the same over a long period of time.

Step 3: Recompute Prices after an interval. We then find out the prices of the SAME bundle of goods, using the SAME methodology for finding prices after some time. Generally, this is done on a monthly or weekly basis. Below we give data over a one-year interval.

| | | Oct-18 | Oct-19 |
|-----------------------------|---------|--------|--------|
| Item | QTY | Price | Price |
| Wheat Flour, Bag | 10 kg | 182.8 | 198.9 |
| Rice Basmati Broken | 1 kg | 37.8 | 44.6 |
| Chicken Farm, Broiler, Live | 1 kg | 83.4 | 90.1 |
| Milk, Fresh, Unboiled | 1 Ltr | 30.5 | 35.4 |
| Cooking Oil, Tin, (SN) | 2.5 Ltr | 316.3 | 306.4 |
| Pulse Masoor, Washed | 1 kg | 71.4 | 88.9 |
| Potatoes | 1 kg | 15.2 | 17.8 |
| Sugar, Refined | 1 kg | 27.9 | 25.5 |
| Tea Prepared, Average Hotel | Cup | 6.9 | 8.9 |

| | | | |
|-------------------------------------|-------|------|------|
| Electricity Charges (Average) | Unit | 4.4 | 5.3 |
| Petrol, Super | 1 Ltr | 57.8 | 55.8 |

The table gives the prices of the same unit on 25 October 2018 and on 25 Oct 2019, one year later. The changes in the price measure the “inflation” in that good. Note that there was disinflation, reduction in prices, in sugar, and in Petrol, for reasons known to consumers in Pakistan.

Step 4: Compute Inflation in each category separately. Given the two prices, separated by one year of time, we can compute the annual inflation rate for each commodity separately. This is done in the table below:

| Item | QTY | Oct-18 | Oct-19 | Inflation |
|-----------------------------|---------|--------|--------|-----------|
| Wheat Flour, Bag | 10 kg | 182.8 | 198.9 | 8.80% |
| Rice Basmati Broken | 1 kg | 37.8 | 44.6 | 18.08% |
| Chicken Farm, Broiler, Live | 1 kg | 83.4 | 90.1 | 8.05% |
| Milk, Fresh, Unboiled | 1 Ltr | 30.5 | 35.4 | 16.26% |
| Cooking Oil, Tin, (SN) | 2.5 Ltr | 316.3 | 306.4 | -3.14% |
| Pulse Masoor, Washed | 1 kg | 71.4 | 88.9 | 24.49% |
| Potatoes | 1 kg | 15.2 | 17.8 | 16.95% |
| Sugar, Refined | 1 kg | 27.9 | 25.5 | -8.67% |
| Tea, Avg Hotel | Cup | 6.9 | 8.9 | 28.80% |
| Electricity Charges (Avg) | Unit | 4.4 | 5.3 | 21.56% |
| Petrol, Super | 1 Ltr | 57.8 | 55.8 | -3.51% |

How do we compute this inflation number? The standard formula for Inflation is:

$$\text{Inflation} = (\text{Current Price} - \text{Previous Price})/\text{Previous Price}$$
. This is also equal to the ratio of current price to previous price minus 1. This means the BASE is Oct 2018 Price. This is a CONVENTION. We can also make the base equal to the current price, or the average price over

the two years, or many other possibilities. As long as we stick to one method, it usually does not matter very much.

Step 5: Assign WEIGHTS to each commodity. It does not make sense to take the simple average of all of the 11 inflation rates, and call it the overall inflation. This is because not all commodities are equally important. How can we assess the relative importance of the different commodities? Actually, there are many different sensible ways of doing this. The one we choose is not because it is the best. Rather, the method chosen is the most convenient to use. PBS uses the volume of TOTAL SALES of the commodity in the Fiscal Year 2007-8 as the weight for that commodity. Ideally, the total sales in the current year should be used, But data on this is not easily available and is expensive to gather. That is why the BASE of the price index is calculated once, and then changed after long intervals. Currently, PBS is in process of shifting the base of the price index from 2007-8 to 2015. This involves calculating the weights for all of the commodities in the price index, by calculating the total volume of sales for that commodity in Fiscal Year 2015.

The table below gives ARTIFICIAL numbers for total sales for the chosen 11 SPI commodities. To illustrate the process of calculating weights. We can take these numbers as sales in Millions of PKR.

| | | FY 2007 | | |
|------------------------|-----------|---------|---------|---------|
| Item | Inflation | Sales | Percent | Product |
| Wheat Flour, Bag | 8.80% | 10.9 | 18.99% | 1.67% |
| Rice Basmati Broken | 18.08% | 1.9 | 3.32% | 0.60% |
| Chicken Broiler, Live | 8.05% | 3.6 | 6.20% | 0.50% |
| Milk, Fresh, Unboiled | 16.26% | 16.8 | 29.34% | 4.77% |
| Cooking Oil, Tin, (SN) | -3.14% | 2.3 | 4.00% | -0.13% |
| Pulse Masoor, Washed | 24.49% | 0.5 | 0.85% | 0.21% |
| Potatoes | 16.95% | 1.3 | 2.18% | 0.37% |
| Sugar, Refined | -8.67% | 2.7 | 4.76% | -0.41% |
| Tea Prep, Avg Hotel | 28.80% | 0.8 | 1.38% | 0.40% |
| Electricity Avg | 21.56% | 11.5 | 20.06% | 4.32% |
| Petrol, Super | -3.51% | 5.1 | 8.92% | -0.31% |
| | Total= | 57.4 | 100% | 11.99% |

The third column of numbers gives the percentage of sales, as a proportion of the total sales of all 11 commodities. Sales of Wheat flour were 10.9 Million PKR, while Total Sales for all 11 commodities were 57.4 Million PKR. So the proportion of Wheat was $10.9/57.4 = 18.99\%$. Similarly, we can calculate the percentage weight of each of the commodities by looking at how much money was spent on that commodity, as a percentage of the total expenditure.

Step 6: Compute the WEIGHTED Average of Inflation rates in each category. This consists of two steps. The first step is to Multiply Inflation for each commodity by the PROPORTION of the COMMODITY in the OVERALL Budget for the Nation. This is listed in the last column of numbers above. For Wheat, we have 8.8% Inflation multiplied by the 18.99% Share gives a 1.67% contribution to inflation. The inflation rate for each of the commodities is multiplied by its percentage weight to get the contribution. The largest contribution to inflation comes from 4.77% in Milk and 4.33% in Electricity. At the same time, Sugar and Electricity contribute negatively and actually bring down inflation. ADD up all of the shares to get total inflation of 11.99% in the SPI. This completes the technical details of how inflation is computed from the SPI – the sensitive price index. This number is supposed to measure increases in the Cost of Living for the Poor, in terms of essential commodities. But it does not include expenses on rent, education, health, which are actually very important. So it is a very imperfect index, and there is substantial room for creating better measures of how expensive life is for the poor. We will discuss some alternatives in later lectures.

Concluding Remarks: It seems that numbers are objective, but our goal has been to show that there is a lot of subjectivity in the construction of the index numbers. In particular, the following aspects have some subjectivity involved:

1. Choice of Commodities has SOME flexibility – not arbitrary, but not fixed
2. Finding the Prices has some flexibility
3. Assigning Weights is really important, and is NOT done well. We are using 2007-8 weights, which may have changed a lot.

I have NOT explained the method of calculation in the standard way. I just chose the easiest way to understand. OTHER methods for the same calculation will be discussed later.

4C: Composite Commodities: Laspeyres and Paasche Indices

This lecture provides an alternative approach to computing price index. First we review the Method for computing inflation from the Previous Lecture (bit.ly/dsia04c):

| | Oct-2018 | | Oct-2019 | |
|---------------------|----------|-------|----------|-----------|
| Item | QTY | Price | Price | Inflation |
| Wheat Flour, Bag | 10 kg | 182.8 | 198.9 | 8.80% |
| Rice Basmati Broken | 1 kg | 37.8 | 44.6 | 18.08% |
| Chicken Farm, Live | 1 kg | 83.4 | 90.1 | 8.05% |

| | | | | |
|-------------------------|---------|-------|-------|--------|
| Milk, Fresh, Unboiled | 1 Ltr | 30.5 | 35.4 | 16.26% |
| Cooking Oil, Tin, (SN) | 2.5 Ltr | 316.3 | 306.4 | -3.14% |
| Pulse Masoor, Washed | 1 kg | 71.4 | 88.9 | 24.49% |
| Potatoes | 1 kg | 15.2 | 17.8 | 16.95% |
| Sugar, Refined | 1 kg | 27.9 | 25.5 | -8.67% |
| Tea Prepared, Avg Hotel | Cup | 6.9 | 8.9 | 28.80% |
| Electricity (Average) | Unit | 4.4 | 5.3 | 21.56% |
| Petrol, Super | 1 Ltr | 57.8 | 55.8 | -3.51% |

We calculate Inflation for EACH good separately. Then we take a WEIGHTED average. The Weights are the PROPORTION of MONEY SPENT on the COMMODITY in the Fiscal Year 2008. These are the Aggregate Consumption numbers for the Nation as a whole. Inflation is relatively easy for ONE commodity – we just measure how much the price changes. It is difficult for Multiple Commodities because the price of each commodity changes in a different way. In the above table, the highest rate of inflation is 28.8% for Tea, and the lowest is -3.5% for Petrol – so how can we find ONE number to represent this entire range of price changes?

The standard method is to use a WEIGHTED Average to get to ONE number for inflation. The WEIGHTS are the MONEY SPENT on the commodity – this captures the IMPORTANCE of the commodity within budget. But there are many problems with this procedure.

PROBLEM 1: These weights keep changing with time. We used FY 2008 because data was gathered for this purpose. Change of BASE for price index is currently underway to update to 2015. But this is still behind current time patterns of consumption. How much sense does it make to use consumption patterns of 2008 to compute inflation rates in 2019?

PROBLEM 2: These are AGGREGATE weights for the nation as a whole. These may not be representative of purchasing patterns of individuals, especially of the poor subgroups. This point will be made clearer in this lecture, which provides an **Alternative Method for Computing Price Index and Inflation**

We are reproducing a table from the previous lecture (BIT.LY/DSIA04C) which shows how we compute weights for the different commodities:

| | | 10/18 | FY 2008 | Composite |
|------|-----|-------|---------|-----------|
| Item | QTY | Price | Sales | QTY |

| | | | | |
|-----------------------|---------|-------|---------|---------|
| Wheat Flour, Bag | 10 kg | 182.8 | 10901.4 | 59.63 |
| Rice Basmati Broken | 1 kg | 37.8 | 1903.2 | 50.39 |
| Chicken Broiler, Live | 1 kg | 83.4 | 3558.8 | 42.68 |
| Milk, Fresh, Unboiled | 1 Ltr | 30.5 | 16836.8 | 552.93 |
| Cooking Oil, Tin | 2.5 Ltr | 316.3 | 2295.3 | 7.26 |
| Pulse Masoor | 1 kg | 71.4 | 489.2 | 6.85 |
| Potatoes | 1 kg | 15.2 | 1250.1 | 82.14 |
| Sugar, Refined | 1 kg | 27.9 | 2734.0 | 97.92 |
| Tea Avg Hotel | Cup | 6.9 | 791.3 | 114.52 |
| Electricity (Avg) | Unit | 4.4 | 11512.8 | 2640.55 |
| Petrol, Super | 1 Ltr | 57.8 | 5118.8 | 88.51 |

The weights are computed from the BUDGET shares, the amount of money spent on each commodity. We have data on the Price of these commodities on 25 Oct 2018. The table above lists (hypothetical) MONEY SPENT on total sales of the commodity in FY 2008, this time in Thousands of PKR – Previous Table was in Millions of PKR. Since Price x QTY = Money Spent, we can calculate TOTAL QTY Purchased. This is given in the Last Column. The Quantity Purchased is obtained by dividing the Money Spent by the Price per unit. The last column can be thought of as a COMPOSITE GOOD. If we think of the ENTIRE nation as ONE household, then the data for 2008 shows that this nation purchased 59.63 thousand 10 Kg bags of Wheat, 50.39 thousand Kg of Rice, and so on. If we DEFINE consumption to be this FIXED basket of goods purchased in exactly these proportions, this basket of goods is called a composite good. Then, the Price Index is just the Price of the Composite Good. Also, Inflation just measures Increases in the Price of CG. By creating a composite good, we have assembled all the different goods into one package, which allows us to measure inflation by pricing this package across time.

The standard method for creating a composite good is to use commodity bundles purchased by the whole nation for a fixed Base Year. Until recently, the Pakistan Bureau of Statistics was using FY 2008 as the base year, and now it has recently been changed to 2015. This method is called the Laspayre's Price Index. This method fixes consumption patterns, and the composite good, in some past year. The advantage of this method is that we only need to gather detailed micro-data on patterns of consumption for ONE year. After that, the pattern of consumption remains fixed, avoiding the need for costly surveys to find current consumption patterns. The disadvantage is that the weights are not representative of the current proportions of the good in the consumption bundle today. We are using consumption patterns of 2008 to

calculate inflation in 2018. A superior alternative is the Paasche Index. This method uses the Composite Good based on CURRENT consumption patterns. This is rarely used in practice because the information on current consumption bundles required to compute this index is rarely available.

We will now illustrate all of these ideas by a very specific, simple, and artificial example. A Price Index is just the Price of a Composite Good. A Composite Good is just A bundle of commodities, regarded as ONE good. To illustrate this consider FOOD as a composite good, which is made up of different quantities of Wheat, Rice, Milk, and Chicken. The table below uses 30 Kg of Wheat, 10 kg of Rice, 50 Liters of Milk, and 10 Kg of Chicken as the composite good FOOD:

| | Qty | Price 2018 | Budget 2018 | Price 2019 | Budget 2019 |
|---------|-----|------------|-------------|------------|-------------|
| Wheat | 30 | 180 | 5400 | 198 | 5940 |
| Rice | 10 | 37 | 370 | 45 | 450 |
| Milk | 50 | 30 | 1500 | 35 | 1750 |
| Chicken | 10 | 80 | 800 | 88 | 880 |
| | | | 8070 | | 9020 |

Using 2018 price, we calculate the amount spent on each commodity and add them up to get the total budget required to purchase FOOD (the composite commodity). This is 8070 PKR in 2018 and increases to 9020 in 2019. Then we can calculate the inflation rate as $11.77\% = (9020/8070 - 1)$.

While this gives a clear answer to the question of how to compute inflation, there remains the question of ‘ How to choose the FOOD bundle?’ What are the quantities we should use for each of these commodities? As we will show this choice matters a lot in computing the price index and the inflation rate. The Conventional Choice is the Laspeyres Index. That is, we fix ONE YEAR as the BASE for the price index. We measure AGGREGATE CONSUMPTION of the commodities in our bundle for the entire nation. Instead of treating the entire nation as ONE household, a more sensible alternative is to look at each household separately. This leads to much greater insight regarding the meaning of price index and inflation, as we now show.

The table below considers the consumption patterns of one (hypothetical) household in the two years 2018 and 2019:

| | Q 2018 | P 2018 | B 2018 | Q 2019 | P 2019 | B 2019 |
|-------|--------|--------|--------|--------|--------|--------|
| Wheat | 365 | 180 | 65700 | 400 | 198 | 79200 |

| | | | | | | |
|---------|-----|----|--------|-----|----|--------|
| Rice | 220 | 37 | 8140 | 200 | 45 | 9000 |
| Milk | 450 | 30 | 13500 | 550 | 35 | 19250 |
| Chicken | 120 | 80 | 9600 | 150 | 88 | 13200 |
| Pulse | 300 | 20 | 6000 | 250 | 40 | 10000 |
| | | | 102940 | | | 130650 |

For this Household, the first column measures the Quantity for each of the FIVE goods, Wheat, Rice, Milk, Chicken, and Pulse, that the household ACTUALLY purchased over the entire year 2018. The second column gives the PRICES at which these commodities were purchased. Of course, this price fluctuates over the year. For the purpose of computing the budget, we actually need the average price at which the household purchased these goods over the entire year. Let us assume that the Oct 2018 price we have is representative of this average price. Then the third column gives us the FOOD budget of this household in 2018. Now, these same numbers are replicated for the year 2019 in the last three columns. The purchasing pattern of the household changes, with increases in Wheat, Milk, and Chicken purchases, and decreases in Rice, and Pulse. The budget spent on food increases to 130,650 from 102,940. Does this increase of 26.9% ($=130,650/102,940 - 1$) represent INFLATION? NO – because the Composite Commodity CHANGED. We must keep Quantities FIXED to measure inflation. Part of the budget increase comes from increases (or changes) in purchased QUANTITIES. Another part comes from changes in prices. The figures above MIX both of these effects, due to price change and due to quantity change. We must keep Quantity FIXED in order to measure the effect of price change only – the inflation – on the budget. We have TWO CHOICES. Either we can keep quantities fixed at the PREVIOUS year 2018 levels – the Laspeyre method. Or we can use the Paasche method, which keeps quantities fixed at the CURRENT YEAR numbers. We now illustrate both of these methods.

The Laspeyres method uses Previous Year Prices as the Base. The actual quantities purchased in 2019 are replaced by the previous year's quantities, highlighted in red. This change in quantities leads to a change in the BUDGET for 2019, which is also shown in red.

| | Q 2018 | P 2018 | B 2018 | Q 2018 | P 2019 | B 2019 |
|---------|--------|--------|--------|------------|--------|--------------|
| Wheat | 365 | 180 | 65700 | 365 | 198 | 72270 |
| Rice | 220 | 37 | 8140 | 220 | 45 | 9900 |
| Milk | 450 | 30 | 13500 | 450 | 35 | 15750 |
| Chicken | 120 | 80 | 9600 | 120 | 88 | 10560 |
| Pulse | 300 | 20 | 6000 | 300 | 40 | 12000 |

| | | | | | |
|--|--|--|--------|--|---------------|
| | | | 102940 | | 120480 |
|--|--|--|--------|--|---------------|

If the household did not change its consumption pattern, and purchased exactly the same goods in 2019 that it did in 2018, then its budget would have been 120,480. Now, we can calculate inflation as $120480/102940 - 1 = 17.04\%$. This is the Correct measure of inflation for THIS Household, using Laspayre's Index. Note that for a correct computation of the budget, PRICES should be averaged over the entire year, NOT fixed at one point in time like Oct 2019. This is a minor issue that we will ignore in the present lecture.

The alternative to Laspayre is to use Paasche, which uses the CURRENT year as the base. This calculation is shown below:

| | Q 2019 | P 2018 | B 2018 | Q 2019 | P 2019 | B 2019 |
|---------|---------------|--------|---------------|--------|--------|--------|
| Wheat | 400 | 180 | 72000 | 400 | 198 | 79200 |
| Rice | 200 | 37 | 7400 | 200 | 45 | 9000 |
| Milk | 550 | 30 | 16500 | 550 | 35 | 19250 |
| Chicken | 150 | 80 | 12000 | 150 | 88 | 13200 |
| Pulse | 250 | 20 | 5000 | 250 | 40 | 10000 |
| | | | 112900 | | | 130650 |

Instead of using the actual quantities purchased by this household last year, we use the CURRENT year purchases, and price them at last year's prices. The first column of Q 2018 (actual purchases in 2018) is replaced by Q 2019, the actual purchases in 2019, as shown highlighted in red. This leads to a change in the BUDGET for 2018, again highlighted in red. We can now compute the Paasche measure of inflation as $130650/112900 - 1 = 15.72\%$. This is rather different from the Laspayre measure of 17.04% computed earlier.

There are many OTHER options. Instead of taking either of the two years as the base, we could take the average amount of purchases in both years, OR just average Paasche and Laspeyres index. The important question is: "Is there a CORRECT measure of inflation?:" The answer is NO: We should think of inflation as a QUALITATIVE phenomenon, which we are trying to measure imperfectly with numbers. This means that A RANGE of numbers can be suitable, and no one measure can adequately describe inflation.

To see how the choice of commodity bundle matters, we will show that different households can have VERY DIFFERENT measures of inflation. Here we compare two different

households. One of them eats only Rice and Pulses R&P (Dal & Chawal in URDU). We look at the inflation rate using the budget of this household in the table below:

| | R&P | P 2018 | B 2018 | R&P | P 2019 | B 2019 |
|---------|-----|--------|--------|-----|--------|--------|
| Wheat | 0 | 180 | 0 | 0 | 198 | 0 |
| Rice | 500 | 37 | 18500 | 500 | 45 | 22500 |
| Milk | 0 | 30 | 0 | 0 | 35 | 0 |
| Chicken | 0 | 80 | 0 | 0 | 88 | 0 |
| Pulse | 500 | 20 | 10000 | 500 | 40 | 20000 |
| | | | 28500 | | | 42500 |

For this R&P household, the Inflation rate is very high at 49.1%. This is because the price of pulse (Daal) has doubled from 20 PKR to 40 PKR, and this household spends a lot of money on Pulses. So its budget is very badly affected by inflation, going from 28,500 to 43,500 for nearly 50% inflation rate.

Another household has very different consumption patterns, eating only Wheat, Milk, and Chicken – WMC:

| | WMC | P 2018 | B 2018 | WMC | P 2019 | B 2019 |
|---------|-----|--------|--------|-----|--------|--------|
| Wheat | 500 | 180 | 90000 | 500 | 198 | 99000 |
| Rice | 0 | 37 | 0 | 0 | 45 | 0 |
| Milk | 500 | 30 | 15000 | 500 | 35 | 17500 |
| Chicken | 500 | 80 | 40000 | 500 | 88 | 44000 |
| Pulse | 0 | 20 | 0 | 0 | 40 | 0 |
| | | | 145000 | | | 160500 |

Because the prices of these commodities did not rise by too much, the budget of the WMC household increases only by 10.7%, going from 145,000 to 160,500.

Since different households will face different inflation rates, depending on their purchase patterns, we can ask what is the range of variation for this inflation rate? How much can it vary? To answer this question, we must look at the inflation rate for each commodity separately, as in the table below

| | P 2018 | P 2019 | Inflation |
|---------|--------|--------|-----------|
| Wheat | 180 | 198 | 10.0% |
| Rice | 37 | 45 | 21.6% |
| Milk | 30 | 35 | 16.7% |
| Chicken | 80 | 88 | 10.0% |
| Pulse | 20 | 40 | 100.0% |

This shows the inflation rates for each of the five commodities in the food bundle separately. The range of inflation rates is between 10% and 100%. A Weighted Average can vary between max 100% and min 10%. If a household purchases a lot of Pulses, then it will see inflation near 100%. If another household purchases mostly Wheat and Chicken, it will see the minimum possible Inflation of 10%. ALL combinations of weights will come out BETWEEN these two numbers. This is because a weighted average is always within the range of numbers that are being averaged.

Concluding Remarks

1. Inflation varies by Household, according to their purchase patterns.
2. Purchase pattern changes with time, so no single number measures inflation.
3. Making ARBITRARY conventions allows us to ASSIGN a number, but this number is not an ACCURATE measure of inflation. It should be taken as an INDICATOR of qualitative characteristics. A range of numbers may be suitable to describe inflation, even for a single household.
4. No single number describes inflation for the country as a whole. The idea that taking the AGGREGATE consumption bundle in 2008 as the composite commodity leads to accurate measures – but this is WRONG. This involves reducing the country to one household and ignoring diversity in consumption patterns across different income groups as well as regional groups.
5. It is possible to get MORE accurate measures by STRATIFYING the population according to consumption patterns. Then we can assign different (approximate) inflation numbers to different subgroups of the population.
6. The variations in consumption patterns are also ENDOGENOUS and CAN BE CHANGED by campaigns. Endogenous means that prices affect these patterns – if something becomes expensive, less use will be made of it. Also, tastes can be changed by many different factors, leading to changes in consumption patterns. Such changes can also affect the inflation rate.

4D Big Data: Many Inflations

We saw in the previous lecture that Inflation rates vary by household. The rate depends on what goods are purchased by the household. Even within one household, the rate of inflation varies with the base – Laspeyres takes last year as the base, while Paashce takes current year as the base – both of them yield different numbers. In this lecture, we consider whether we can just look at ALL of the inflation rates together, instead of trying to reduce them to one number.

The mindset of statistics is a reflection of Sir Ronald Fisher's statement that Statistics is about the reduction of data. In his era, it was impossible to directly analyze large data sets, and the only hope for analysis was to reduce large data sets to a small and manageable size. This is exactly the opposite of the Modern Mindset: we would like to analyze BIG Data Sets because they contain a lot more information, and WE CAN analyze them using currently available computer capabilities.

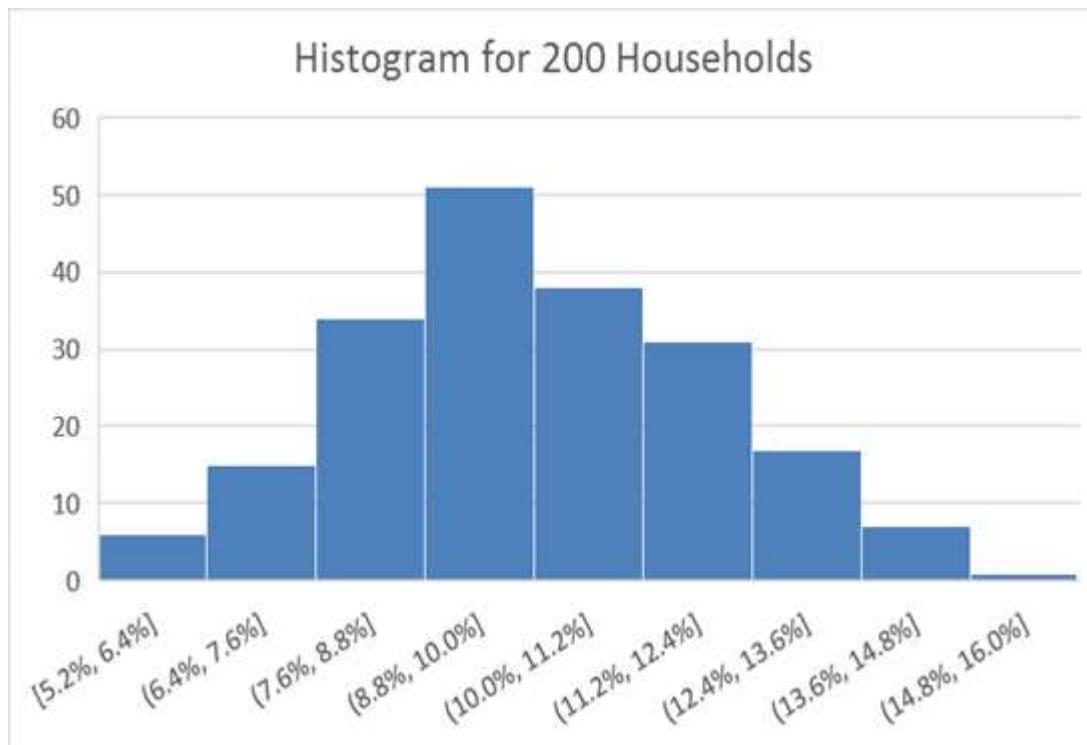
We have seen that each commodity bundle leads to a separate inflation rate. Can we reduce thousands – millions – of inflation numbers to ONE number? NO – not without serious loss of information. Conventional Statistics looks for Sufficient Statistics – a small set of numbers that summarizes ALL of the information in the data. These exist if data density follows restrictive ASSUMPTIONS. GENERALLY speaking, it was ASSUMED that all data is NORMALLY distributed. When these assumptions are approximately valid, they allow us to reduce the data and enable analysis in a pre-computer era. Typically, these assumptions are FALSE, especially in BIG data sets. Because of advances in technology, it is NOW possible to DIRECTLY analyze big data sets, WITHOUT trying to reduce them. We need to learn NEW WAYS of thinking, and NEW techniques of analysis for large data sets. In this lecture, we will explain how to analyze Inflation for EACH household separately.

Actual analysis of real data sets can be done using surveys. For example, the Household Income and Expenditure Survey (HIES) in 2006 took a sample of 39,677 HH's in Pakistan. Even though 40,000 is a big number, it is a “small” sample chosen as a representative of the entire population of around 180 million at that time. It is now BECOMING possible to track each individual separately – China, USA, and Europe now have the technology available to keep track of all consumer purchases over one year for each household separately. However, this data is not currently collected in this format. The data requirements for tracking annual consumption bundles are LARGE. They are much lower for PRICES which tend to remain stable over time and are the same for all individuals. That is why the Laspayre index is typically used, because you track consumption patterns only for one year, and then use that as representative for all years, while measuring price changes. Since real data of the type we need is not available, we will generate an artificial data set to illustrate the concepts. The GOAL is to understand what inflation numbers mean.

As we have seen, Inflation numbers vary by Household, and no ONE number represents this data. To illustrate this concept, consider an artificial data set for 10 households:

We consider 8 Essential Food Commodities listed above. The actual price inflation rates for each commodity going from 2018 to 2019 are listed in the first column. Different Households purchase different amounts of these commodities. Randomly chosen amounts ranging from 5 to 95 are listed under each household. This gives us consumption bundles for 10 households. We can now compute the inflation rate for each household. Multiply inflation rates by the units of quantity and sum, and divide this by the sum of the quantities. Assuming the data is entered in an EXCEL spreadsheet starting from A1, the formula which gives the inflation rate for HH 1 would be entered in cell C10 as: =SUMPRODUCT(B2:B9,C2:C9)/SUM(C2:C9). Similarly, we can compute inflation for each of the households separately. Pulse, Rice, and Milk have high inflation. So households that purchased more of these commodities would have high inflation numbers. Sugar and Oil have negative inflation, while Wheat and Chicken have the lowest positive inflation. So households that have more of these four goods in their consumption bundle would experience lower inflation. For the random data chosen above, the range of inflation is from 8.08% to 11.68%. No one number can summarize this data set. For 10 numbers, we can look at them directly and understand the whole data set. When the data set is larger, this is not possible. We examine this case next.

We generate Random Consumption Patterns for 200 HH & compute 200 Inflation numbers. These vary from a low of 5.2% to a high of 16.06%. Our minds are not built to process 200 numbers directly. Statistics are graphical aids to translate data into visual forms which can easily be understood. A Histogram gives a picture of the 200 inflation numbers:



From the histogram, we can learn a lot more about inflation for the 200 households. The Histogram Data can be given in a tabular form as follows:

| | | |
|--------|----|----|
| 5.20% | 6 | 1 |
| 6.40% | 15 | 3 |
| 7.60% | 32 | 6 |
| 8.80% | 51 | 10 |
| 10.00% | 40 | 8 |
| 11.20% | 30 | 6 |
| 12.40% | 17 | 3 |
| 13.60% | 8 | 2 |
| 14.80% | 1 | 1 |

Each bin is represented by its lowest point. The first bin which goes from 5.2% to 6.4% has 6 HHs. The Modal Bin is the one with the largest number of households. There are 55 HH in the range of 8.8% to 10.0%, so that this is the Mode for the data. We also note that the three central bins, going from 7.6% to 11.2% contain 123 HHs with is 61.5% of the total 200 HHs. In data reduction, we aim to find a smaller data set having nearly the same distribution as the larger one. One possibility is given in the third column of numbers above, which replaces each of the 5 HHs with one HH in the same category. The last column REDUCES data from 200 to 01 HHs having NEARLY the same distribution.

Classical methods of data reduction rely on the measure of the center of the distribution as a Representation of the data. In this context, let us examine the Mode & Mean of these 200 inflation numbers. We can calculate that the Mean = Average = 10.0%. Does this number provide a good **representation** of the data: Does 10% represent the experience of many HHs? It is obvious that the answer is NO. There is a Technical Formula related Mean to Data: Add up all the numbers and divide by the total count (200 in this case). How does this number relate to the data set? The Meaning comes from a Theoretical Assumption. If Data follow a Normal Distribution, then the average of the data is the best estimate of the mean of this normal distribution. Other interpretations can be made under different types of theoretical assumptions. These are always complex and depend on the validity of the assumptions.

The key feature which distinguishes Descriptive Statistics from Conventional Theoretical Statistics is the attempt to directly understand the data, without theoretical assumptions. The average of the data does not have a clear meaning without theoretical assumption. A more directly comprehensible central value is the MODE of the data. This is the bin containing the largest amount of data. As already indicated, the Mode is the bin [8.8%, 10.0%]. The largest number of families (51/200) saw inflation in this range. This is a somewhat MISLEADING statement.

Compared to WHAT? We should ALWAYS ask this question. Here we are comparing to the OTHER BINS – but this is not mentioned in the statement, leading to possible confusion. If we want to capture the experience of most families, meaning the majority of them, the three central bins contain 123 families or 61.5% of all families. Thus we can say that most families (123/100) saw inflation rates between 7.6% to 11.2%. So if we want to REPRESENT the experience of most families, this range [7.6%,11.2%] is good. A better way to do this is to look at the inter-quartile range, discussed later in this lecture.

The median and the average attempt to provide one number as a REPRESENTATION of the data. Another objective of a Central Measure is to provide a benchmark. This is best done by the MEDIAN. For this data set, MEDIAN Inflation = 9.9%. This means that

Half of the households had inflation rates below 9.9%. The other half experienced inflation at rates above 9.9%. This 9.9% is a benchmark. HHs which experienced inflation above 9.9% experience HIGHER (than median) inflation, those below the benchmark experience LOWER (than median) inflation. This divides the HHs into two equal groups of LOW and HIGH. How do you compute the median?

1. SORT the data from highest to lowest.
2. Even data count (200): Median is [100th, 101th] data points = [9.93%,9.94%]
3. Odd data count: Middle Value exactly.

A Technical Definition if the median is that it satisfies the following two conditions:

1. $\geq 50\%$ of the data should be \leq MEDIAN
2. $\geq 50\%$ of the data should be \geq MEDIAN

Concluding Remarks

1. We have seen that the nature of Inflation is such that it is variable across HHs, and cannot be captured by ONE number. We have used histograms, modal bins, and median to try to describe some aspects of the larger data set.
2. The use of one number leads to a loss of credibility because it does not match experience. To improve matters, we need to educate the public that inflation measures average price increase for all goods, and a few goods with spectacular price increases do not capture the average of everything. ALSO, we need to provide a range of values that captures the general experience more adequately than a single number can.
3. Instead of using the average as a central value, the goal of REPRESENTATION can be better achieved by the use of MODAL values in LARGE BINS. Large bins capture more of the data and hence do a better job of representation. Small Bins lead to MISLEADING MODES, as discussed earlier.
4. In this context, the INTERQUARTILE RANGE provides a useful way to pick out the central half of the data. For the 200 numbers in the inflation data set, the MIDDLE HALF

of the data is [8.66%, 11.41%]. This is obtained by sorting the data, and then looking at the range from the 50th to 150th HALF of the households (100/200) experiencing inflation in this range. 25% saw inflation below 8.66% and 25% saw inflation above 11.41%

5. The standard data summaries are Mean and Standard Deviation. These are useful IF Normality Assumption about data distribution holds, and USELESS otherwise. In many situations, small deviations from Normality can lead to very poor performance of the Mean as a data summary. In contrast, the MEDIAN is very robust, and works well for all data distributions.

4E Inflation as a Macroeconomic Concept

One of the CENTRAL concepts underlying this course is the idea that theory and practice cannot be separated. So far we have considered the measurement of inflation in the microeconomic context of how rising prices affect the budgets of households. In this and the next part, we consider the concept of inflation, as it arises in economic theory, from a macroeconomic perspective. As we will see, the calculations we need, and the definitions, change according to the real-world context and purpose.

So far, we have considered Inflation as a measure of the effect of changing prices on budgets. In Economic Theory, the concept of Inflation is closely linked to money. The Quantity Theory of Money (QTM) formula states that $MV = PT$. Money times the Velocity of Circulation of Money equals the general price level times the number of transactions (sales/purchases) within the economy. We can explain this formula as follows.

We can classify monetary transactions into two types: transactions that contribute to the GDP and those which do not. In general, sales of products and services produced this year will be part of GDP, but sales of used goods (produced in previous years) or intermediate goods and raw materials, will not be part of the GDP transactions. The details of how to classify transactions are complex, and not needed for our current purposes in this lecture. The Total amount of money USED in GDP transactions over one year can be written as:

$$M^* = P_1 \times T_1 + P_2 \times T_2 + \dots + P_n \times T_n$$

List all of the millions of transactions where a commodity or service is exchanged for money (in a GDP-producing transaction). Then Quantity Sold x Price per unit is the amount of money that is transferred from one party to the other (used) in each transaction. If we sum ALL the transactions, we get M^* , the total amount of money used to purchase GDP goods over the year. This M^* is also the GDP, by definition. M^* will typically be much larger than M , the total amount of money in existence, because money can be used many times over. Define $V =$ Velocity of Money = How many times money is used, on the average (this average will include money never used as well as money used repeatedly). The equation $MV = M^*$ can be used to DEFINE V.

Let $\{Q_1, Q_2, \dots, Q_k\}$ be a COMPOSITE GOOD. – this is the list of all goods sold throughout the year. Note that the same good would be sold in many different transactions. The

list of goods Q_1, \dots, Q_k is much smaller than the number of transactions T_1, \dots, T_n . It is convenient to introduce the **Dot Product Notation**: If $P = (p_1, p_2, \dots, p_n)$ is the Vector of Prices and $Q = (q_1, q_2, \dots, q_n)$ is the Vector of Quantities, the dot product of these two vectors is defined as

$$P \bullet Q = p_1 \times q_1 + p_2 \times q_2 + \dots + p_n \times q_n$$

In EXCEL: SUMPRODUCT(A1:A6,B1:B6) takes the two columns A1:A6 and B1:B6, multiplies the corresponding entries and adds all of these products:

$$\text{SUMPRODUCT}(A1:A6,B1:B6) = A_1 \times B_1 + A_2 \times B_2 + A_3 \times B_3 + A_4 \times B_4 + A_5 \times B_5 + A_6 \times B_6$$

To explain the macroeconomic theory concept of inflation, we first review the concept of a price index. To calculate a Price Index, we first create a composite good. This is the set ALL sales of ALL goods over a one-year period, called the Base Year. The value of this composite good is equal to the GDP for that year, because this is the value of all the goods produced that year. Now we fix the composite good at the base year levels, and find out how much this combination of goods would cost in other years. This will require having information about the prices of each of the commodities in each year. Price information is much easier to get than the value of the entire amount sold throughout the year. The dot product of the prices for each year and the quantities for the base year is the price index. This is the price of the Composite Commodity in each year. We provide a simple illustrative example:

| Item | QTY 2010 | 2010 | 2011 | 2012 | 2013 |
|-------|-------------|------|------|------|------|
| Wheat | 20 | 37 | 53 | 73 | 96 |
| Rice | 15 | 56 | 64 | 74 | 98 |
| Milk | 25 | 45 | 42 | 67 | 64 |
| Beef | 10 | 75 | 86 | 85 | 92 |
| Price | | 3455 | 3930 | 5095 | 5910 |
| INDEX | | 1.00 | 1.36 | 1.59 | 1.90 |

Suppose the Base Year is 2010. We find out the entire annual sales of ALL the commodities produced. Suppose there are only the four commodities listed – Wheat, Rice, Milk, and Beef. The first column gives the annual sale of Wheat – say 20 million Tons, Rice: 15 million tons were sold. Similarly, 25 and 10 million tons of Milk and Beef were sold in 2010. The next column lists the price per unit in 2010. So Wheat costs 37 Thousand Rupees per ton, and so on for the other commodities. The total spending on the four commodities is the SUMPRODUCT(B2:B5, C2:C5) assuming the above numbers are in an EXCEL table starting at A1. Multiplying the QTY by the Price for each commodity and adding up all four entries give the PRICE of 3455 in the base year 2010. This is the GNP in 2010, at CURRENT prices, in

Local Currency Units. Now we repeat this calculation for 2011, using the prices for 2011, but keeping the quantities fixed at the base year levels in 2010. The value of these four commodities in 2011 is SUMPRODUCT(B2:B5, D2:D5) which comes out to be 3930. This is how much the commodity bundle of 2010 would have cost in 2011. It is a measure of how much the prices have changed from 2010 to 2011 – It does not measure GNP in 2011 because it has no information about the production levels in 2011. We can similarly compute the values of the base year quantities in 2012 and 2013. These numbers come out to be 3455 3930 5095 5910, in the second last row of the table marked PRICE. The units of the quantities are arbitrary, and it is only the percentage changes in these prices that matter for the index. So it is the convention to set the base year value to be 100%, and write the price index in the form of percentage points. This can be done by DIVIDING all of the numbers by 3455, the base year value. This gives the index value for the four years as: 100% 136% 159% 190%. These numbers are the Price Index for the four years, and they measure the rate of increase of prices over these years.

We will now look at the data we will use to calculate inflation and assess the validity of the Quantity Theory of Money as an economic hypothesis. As an accounting identity, QTM is true by definition. These two roles of the QTM – as an accounting identity, and as an economic hypothesis, have to be differentiated clearly to avoid confusion. The WDI (World Development Indicators) data set by the World Bank provides series for GDP Current LCU and GDP Constant LCU, and also the GDP Deflator. These are the series we need to construct the relevant macroeconomic inflation series. Below the explanation and names of the series required in the WDI data set are provided for reference:

NY.GDP.MKTP.CN (GDP in Current LCU) GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current local currency.

NY.GDP.MKTP.KN (GDP in Constant LCU) The definition is the same as the previous one, except for the last sentence: “*Data are in constant local currency*”. The change from current price LCU to constant price is achieved by dividing the current price series by the GDP Deflator – which is exactly the price index for production we have discussed previously. This GDP Deflator series is also in the WDI Data Base as:

NY.GDP.DEFL.ZS (GDP Deflator) The GDP implicit deflator is the ratio of GDP in current local currency to GDP in constant local currency. The base year varies by country.

Of the three series, only two are necessary since $\text{GDP}(\text{Current LCU}) = \text{GDP}(\text{Constant LCU}) \times \text{GDP Deflator}$. We can get the third series from any two. The process of calculation involves creating the Current LCU GDP and the GDP Deflator (or Price Index for GDP) and then dividing the GDP Current LCU by the deflator to get the GDP Constant LCU. We illustrate these data series for just one country, Australia, in the WDI data set.

| Australia | Nominal GDP | Real GDP | GDP Deflator |
|-------------|---------------|---------------|---------------|
| CODE=N | NY.GDP.MKTP.C | NY.GDP.MKTP.K | NY.GDP.DEFL.Z |
| 2005 | 9.20899E+11 | 1.12365E+12 | 81.95632788 |
| 2006 | 9.94803E+11 | 1.15778E+12 | 85.92309612 |
| 2007 | 1.08306E+12 | 1.20156E+12 | 90.13759578 |
| 2008 | 1.17595E+12 | 1.2469E+12 | 94.30988396 |
| 2009 | 1.25222E+12 | 1.26393E+12 | 99.07305287 |
| 2010 | 1.29338E+12 | 1.29338E+12 | 100 |
| 2011 | 1.39906E+12 | 1.31803E+12 | 106.15 |

The first column is the GDP in Current LCU for Australia for the years 2005 to 2011. This is also called the Nominal GDP. The second column is the GDP in constant LCU. This is also called the real GDP. The third column is the GDP Deflator, or the price index. Note that the index is 100 in 2010, which shows that 2010 is the base year for this index. Dividing the nominal GDP by the GDP Deflator (in percentages) will give us the real GDP for Australia.

In order to evaluate the QTM as an economic hypothesis, we need to partition economic Growth into two parts: Inflation and Real Growth. As we move from 2010 to 2011 both prices and quantities changed, and increased. We would like to separate the increase into two parts, one due to increase in prices, and the other due to increase in quantities. Here is the data for Australia for these two years:

| | | | |
|---------------|-------------|-------------|--------|
| 2010 | 1.29338E+12 | 1.29338E+12 | 100 |
| 2011 | 1.39906E+12 | 1.31803E+12 | 106.15 |
| Growth | 1.0817 | 1.0191 | 1.0615 |

The second column gives us the Growth in Nominal GDP = 8.17%, calculated as $B2/B1 - 1$. From the last column we get Growth in Prices = 6.15% ($D2/D1-1$). This is the growth rate of prices. Similarly, we can calculate the growth in real GDP to be 1.91% ($C2/C1-1$). So we can conclude that 8.17% can be divided into two parts, 6.15% is due to price increase, and 1.91%. To avoid confusion, it is important to note that: Growth rates are Multiplicative. The nominal GDP in the Base Year is multiplied by the Growth Rate of prices and by the Growth rate of quantities to get the nominal GDP for the next year. This means that $1.0191 \times 1.0615 = 1.0817$ and not $6.15 + 1.91 = 8.06$; growth rate is 8.17% and not 6%.

QTM as an Economic Hypothesis states that growth rates of money affect growth rates of prices, and do not affect growth rates of the real economy. The above three series allow us to compute the growth rates of prices and the real growth rates. To see the relationship between these and money, we need a data series for money. The WDI data set has many series for money, because money can be defined in different ways. Economists talk about Narrow Money (M0), Money (M1), Money+Quasi-Money (M2), and Broad Money (M3). We will not discuss the differences in detail, and we will just choose to work with two of these definitions, to assess the QTM. The series we have chosen from the WDI are listed below, with their names and definitions as given by the World Bank:

FM.LBL.BMNY.CN Broad Money (M3) Broad money (IFS line 35L..ZK) is the sum of currency outside banks; demand deposits other than those of the central government; the time, savings, and foreign currency deposits of resident sectors other than the central government; bank and traveler's checks; and other securities such as certificates of deposit and commercial paper.

FM.LBL.MONY.CN (Money M2) Money is the sum of currency outside banks and demand deposits other than those of the central government. This series, frequently referred to as M1 is a narrower definition of money than M2. Data are in current local currency.

To set up for assessment of the QTM, we need to write the Quantity Equation in Terms of Growth Rates:

$$M(t) \times V(t) = P(t) \times Q(t) \quad \text{and} \quad M(t-1) \times V(t-1) = P(t-1) \times Q(t-1)$$

Take LOGs to convert to ADDITIVE form

$$\log M(t) + \log V(t) = \log P(t) + \log Q(t)$$

$$\log M(t-1) + \log V(t-1) = \log P(t-1) + \log Q(t-1)$$

Subtract the second equation from the first

$\{\log M(t) - \log M(t-1)\} = \log M(t)/M(t-1)$ is approximately growth rate of Money. We will write this as $\%M = \log M(t)/M(t-1)$. Then we can write the above equation as:

$$\log M(t)/M(t-1) + \log V(t)/V(t-1) = \log P(t)/P(t-1) + \log Q(t)/Q(t-1)$$

$$\text{Growth Rates: } \%M + \%V = \%P + \%Q$$

This is Accounting Identity. It DEFINES $V=Velocity$, That is, we can calculate $\%V$ from the above equation, and that will force the equation to be true. However, when we move from Accounting Identity to Economic Theory, the theory may not hold. In the first place, the economic theory says that Velocity is nearly constant – this we can check from the data, and we will find that it is not true. Nonetheless, the theory may still be valid if Velocity is EXOGENOUS: this means that V is not affected by M or P or Q . Learning how to find out about exogeneity is of crucial importance, but not taught in conventional statistics.

The KEY economic hypothesis is that Money affects Prices ONLY, NOT Quantity. This is called Classical Dichotomy: "Money is a Veil", or money is Neutral. It takes TWO forms: STRONG Dichotomy Holds in the short and long run. WEAK Dichotomy: holds Only in Long Run. For the Weak form of the hypothesis, short-run effects of %M on %Q are allowed but these disappear in the long run. This classical economic theory is OPPOSED to the idea of Keynes, who said that: Money is NOT neutral in the short or long run.

In the NEXT lecture, we will look at the data to see whether it supports Keynes or the Chicago school of monetarists who believe in the QTM, using data on GDP and Money from the WDI data set.

4F Failure of Quantity Theory of Money on Australian Data

Dichotomy: holds Only in Long Run. For the Weak form of the hypothesis, short-run effects of %M on %Q are allowed but these disappear in

First we look at the data itself – This is Australian Data, taken from WDI (World Development Indicators) a World Bank data set going from 1960 to 2011 (Updated Data sets are available, going up to 2019).

| Year | V1= GDP /M1 | V2= GDP /M2 | % Real | % rice | %P | % M1 | % M2 |
|------|-------------|-------------|-----------|------------|------------|-----------|------|
| 1960 | 4.67 | 2.24 | | | | | |
| 1961 | 5.07 | 2.26 | 1.0 7% | 1.3 6% | - 1.17% | 2.0 2% | |
| 1962 | 5.03 | 2.12 | 0.6 0% | - 0.10% | 0.8 5% | 3.3 1% | |
| 1963 | 5.07 | 2.08 | 2.6 4% | 0.7 6% | 3.0 7% | 4.3 2% | |
| 1964 | 5.31 | 2.06 | 2.9 5% | 1.4 0% | 2.3 2% | 4.7 6% | |
| 1965 | 5.93 | 2.14 | 2.5 2% | 1.2 8% | - 0.92% | 2.1 9% | |
| 1966 | 5.79 | 2.08 | 1.0 0% | 1.1 5% | 3.1 7% | 3.2 4% | |
| 1967 | 6.03 | 2.13 | 2.6 6% | 2.0 6% | 2.9 2% | 3.7 5% | |

Author Last Name/Book Title

| | | | | | | |
|------------------|------|------|------------|-----------|------------|-----------|
| 19 68 | 6.14 | 2.14 | 2.1 7% | 0.9 7% | 2.4 1% | 2.9 5% |
| 19 69 | 6.28 | 2.19 | 2.9 6% | 2.0 1% | 3.9 5% | 3.9 5% |
| 19 70 | 6.77 | 2.35 | 3.0 2% | 2.1 8% | 1.9 7% | 2.1 7% |
| 19 71 | 7.01 | 2.36 | 1.7 0% | 2.2 0% | 2.3 5% | 3.6 3% |
| 19 72 | 6.45 | 2.18 | 1.6 6% | 2.6 0% | 7.9 2% | 7.8 0% |
| 19 73 | 6.18 | 2.01 | 1.1 4% | 3.7 4% | 6.7 0% | 8.3 9% |
| 19 74 | 7.55 | 2.23 | 1.7 5% | 6.5 9% | - 0.32% | 3.8 1% |
| 19 75 | 7.25 | 2.18 | 0.5 3% | 6.6 5% | 8.9 0% | 8.1 2% |
| 19 76 | 7.79 | 2.27 | 1.1 3% | 5.7 0% | 3.7 1% | 5.0 1% |
| 19 77 | 8.43 | 2.48 | 1.5 2% | 4.7 0% | 2.7 9% | 2.5 0% |
| 19 78 | 8.25 | 2.45 | 0.3 8% | 3.4 4% | 4.7 6% | 4.3 3% |
| 19 79 | 8.09 | 2.48 | 1.7 6% | 3.5 7% | 6.2 1% | 4.7 7% |
| 19 80 | 7.81 | 2.47 | 1.3 1% | 4.1 5% | 6.9 9% | 5.7 0% |
| 19 81 | 8.43 | 2.54 | 1.4 5% | 3.9 6% | 2.0 8% | 4.0 9% |
| 19 82 | 9.75 | 2.65 | 1.3 8% | 4.8 2% | - 0.07% | 4.3 7% |
| 19 83 | 9.10 | 2.52 | - 1.02% | 4.2 3% | 6.1 9% | 5.4 0% |

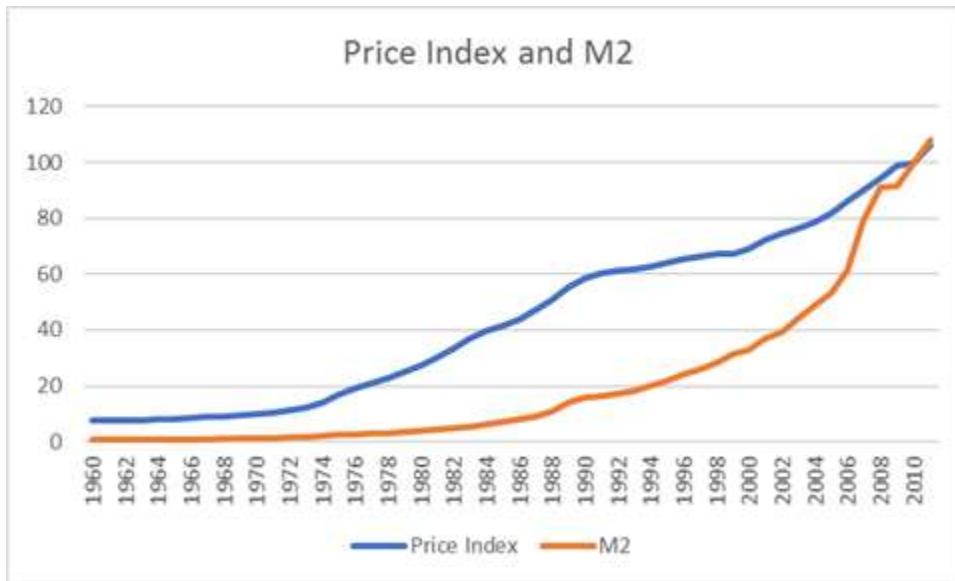
| | | | | | | | |
|-----------|-----------|-----------|------|------------|-----------|-----------|------------|
| 84 | 19 | 9.50 | 2.55 | 2.0 1% | 3.2 8% | 3.4 0% | 4.8 2% |
| 85 | 19 | 10.1 1 | 2.38 | 2.1 5% | 2.0 5% | 1.5 3% | 7.1 6% |
| 86 | 19 | 10.0 3 | 2.40 | 1.9 5% | 2.3 8% | 4.6 8% | 3.9 7% |
| 87 | 19 | 9.15 | 2.27 | 1.1 2% | 2.9 2% | 8.0 3% | 6.4 6% |
| 88 | 19 | 8.97 | 2.19 | 2.3 9% | 3.1 8% | 6.4 1% | 7.1 6% |
| 89 | 19 | 8.45 | 1.90 | 1.6 7% | 3.7 9% | 8.0 4% | 11. 73% |
| 90 | 19 | 8.66 | 1.85 | 1.5 3% | 2.5 9% | 3.0 7% | 5.2 1% |
| 91 | 19 | 8.27 | 1.88 | - 0.15% | 1.3 0% | 3.1 8% | 0.5 1% |
| 92 | 19 | 7.02 | 1.78 | 0.1 9% | 0.6 1% | 7.9 1% | 3.0 9% |
| 93 | 19 | 6.25 | 1.77 | 1.7 5% | 0.3 6% | 7.1 1% | 2.4 2% |
| 94 | 19 | 5.93 | 1.69 | 1.7 2% | 0.4 7% | 4.5 0% | 4.1 2% |
| 95 | 19 | 5.91 | 1.66 | 1.7 0% | 0.9 1% | 2.7 4% | 3.5 5% |
| 96 | 19 | 5.53 | 1.60 | 1.7 0% | 1.1 2% | 5.6 9% | 4.3 9% |
| 97 | 19 | 5.14 | 1.56 | 1.6 6% | 0.5 2% | 5.4 2% | 3.0 6% |
| 98 | 19 | 5.12 | 1.53 | 1.9 2% | 0.5 3% | 2.5 6% | 3.5 2% |
| 99 | 19 | 4.92 | 1.44 | 2.1 0% | 0.2 1% | 4.0 5% | 4.8 1% |

| | | | | | | |
|------------------|------|------|-----------|-----------|-----------|------------|
| 20 00 | 4.80 | 1.48 | 1.6 4% | 1.1 0% | 3.8 0% | 1.5 9% |
| 20 01 | 4.23 | 1.40 | 0.8 2% | 2.0 3% | 8.3 8% | 5.4 0% |
| 2002 | 3.63 | 1.41 | 1.6 7% | 1.1 9% | 9.5 2% | 2.4 1% |
| 20 03 | 3.50 | 1.33 | 1.3 5% | 1.2 2% | 4.1 3% | 5.2 3% |
| 20 04 | 3.52 | 1.28 | 1.7 6% | 1.3 0% | 2.7 7% | 4.6 8% |
| 20 05 | 3.43 | 1.26 | 1.3 6% | 1.6 3% | 4.2 3% | 3.5 9% |
| 20 06 | 3.33 | 1.18 | 1.3 0% | 2.0 5% | 4.5 5% | 6.0 6% |
| 20 07 | 3.21 | 0.99 | 1.6 1% | 2.0 8% | 5.3 0% | 11. 36% |
| 20 08 | 3.29 | 0.94 | 1.6 1% | 1.9 7% | 2.5 7% | 5.7 7% |
| 20 09 | 3.36 | 1.00 | 0.5 9% | 2.1 4% | 1.7 9% | 0.2 1% |
| 20 10 | 3.12 | 0.94 | 1.0 0% | 0.4 0% | 4.6 4% | 3.8 9% |
| 20 11 | 2.97 | 0.95 | 0.8 2% | 2.5 9% | 5.5 5% | 3.3 6% |

The table just illustrates the nature of the data. The main idea of Descriptive Statistics is to teach students how to LOOK at the Data. For this purpose, we PLOT the key data series relevant to the QTM.



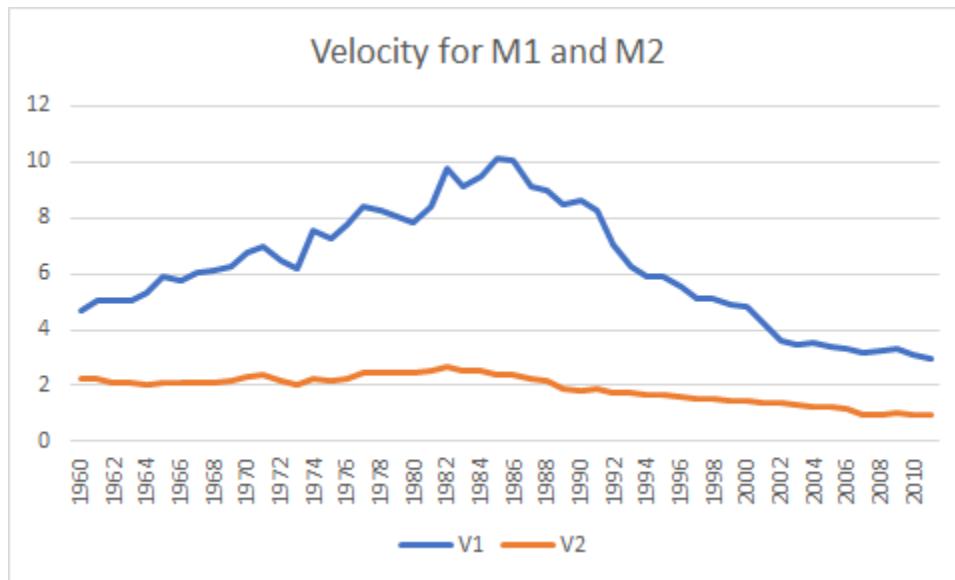
Note that the price index varies from around 10 to 110, while M1 goes from 3.5 Billion AUD to 4 Billion. To put both on the same graph, we need to re-scale M1. We did this by dividing all entries in the M1 column by the value AU\$ 4.15 Billion, the value in 2010. This makes the entry for 2010 equal to 100% just like the price index, and makes the two graphs comparable. Similarly, the graph below shows the behavior of the price level as compared to M2, rescaled to equal 100% in 2010.



These pictures show that the price index and money (M1 and M2) follow rather different patterns. Milton Friedman famously said that: “**Inflation is always and everywhere** a monetary phenomenon”. The previous two graphs show that Friedman is WRONG. In the early portion of the graph, money is increasing slowly, and prices increase rapidly. In a later portion, money rises sharply but prices rise slowly in comparison. From the graph, it is clear that Inflation – rate of change of price index – is NOT solely determined by rate of change of money. The picture is

ENOUGH to show that. Fancy Statistics CANNOT change this basic conclusion. Because of COMPLEXITY of conventional statistics and econometrics, students think that fancy methods might REVEAL some hidden patterns in the data. COMPLEXITY comes from making complex UNTRUE assumptions about the data. By making such assumptions, we can make data say something NOT in the graph.

To further confirm the failure of the QTM as an economic hypothesis (not as an accounting identity), we look at how the VELOCITY behaves. This is the ratio of nominal GDP to Money, and is assumed constant or exogenous. The graph shows the behavior of V1 and V2 the velocity for M1 and M2 respectively:



The Velocity follows different Patterns for M1 and M2. In the case of M1 it increases from 5 to 10 over the period from 1960 to 1986, and then declines from 10 to 3 from 1986 to 2010. The pattern for V2 will be discussed later. This change in velocity shows that there is no stable relationship between Money and Prices – all such relationships depend on the velocity. The idea that Velocity is constant or exogenous means that this factor can be put aside in studying the relationship. But the systematic and strong patterns in Velocity show that we cannot ignore velocity in any serious analysis of relationship between M, P, and Q. Many economists wrote papers on the Breakdown of Money Demand Function in 1980s and 1990s. This corresponds to the change in behavior of V1 around 1989, from increasing to decreasing. In mathematical terms we can write the quantity equation MV = PQ as:

$$\text{Log}(M) + \text{Log}(V) = \text{Log}(P) + \text{Log}(Q)$$

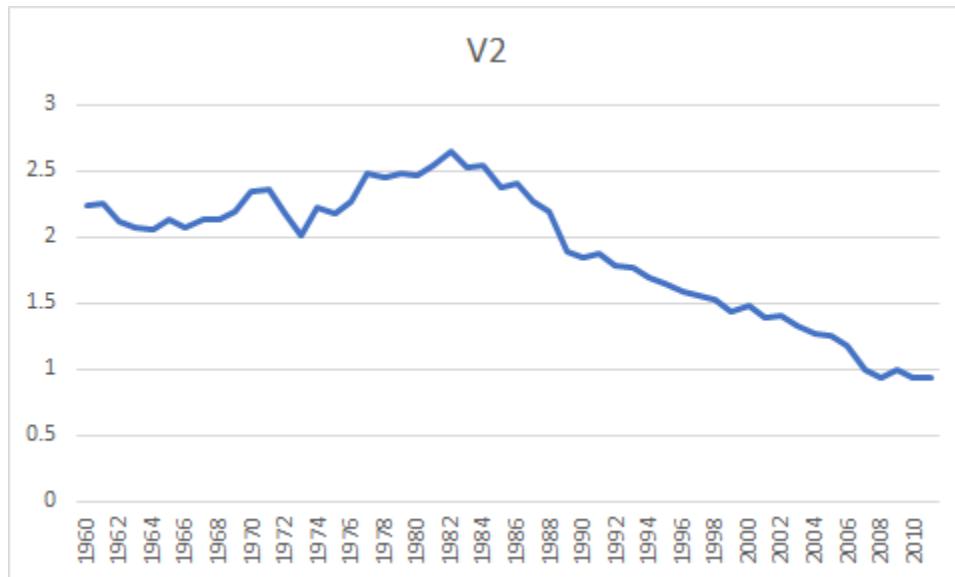
By rearranging terms, we get:

$$\text{Log}(M) - \text{Log}(P) = \text{Log}(Q) - \text{Log} V$$

$$\text{Log}(M/P) = \text{Real Money Demand} = \text{Log}(\text{Real GDP}) - \text{Error}$$

In models like this, the assumption is that the error has no systematic patterns. However, it is clear that the error in the above relationship include the velocity, and this DOES have systematic patterns. This is in violation of the assumptions of the regression model.

The previous graph of V1 and V2 shows that V2 is flatter, showing less variation than V1. This might lead to the hope that the QTM will work for Broad Money, even though it fails for standard Money. To examine this idea, we look at V2, the Velocity for Broad Money on Separate Graph:



There is an important lesson about graphs: Scale Can Create or Hide Patterns. Previously, with both V2 and V1, the Y-axis went from 0 to 12 because V1 had values over 10. Now, with only V2, the scale goes from 0 to 3 only. With this magnification, many details are much more clearly visible than they were on the previous graph. In particular, it is clear that the velocity for M2 is also unstable. From 1960 to 1988 the range of V2 is between 2 and 2.5. In this period, the assumption of constant velocity or exogenous velocity might work. After this period, V2 shows a strong and systematic decline going from 2 to 1.

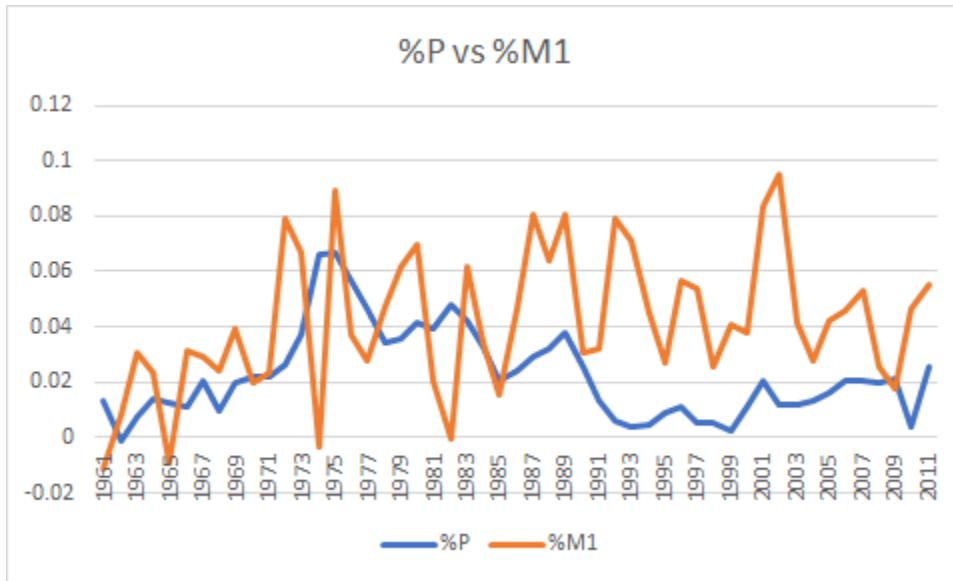
What do we learn from these graphs? We can see that the Velocity behaves in a systematic and predictable way over time. We can see that Velocity is NOT constant. If Velocity is EXOGENOUS, then it MATTERS. It obviously affects the relationship between Money, Prices and Quantity. Any theory which ignores the systematic changes in velocity will be unable to explain the relationship between M, P, and Q.

Next we turn to the ACTUAL assertion of the Quantity theorists. This is that the RATE of increase in money is proportional to the rate of growth of prices. Previous graphs were about the LEVEL of prices and money. As discussed in previous lecture, we measure the growth rates as follows:

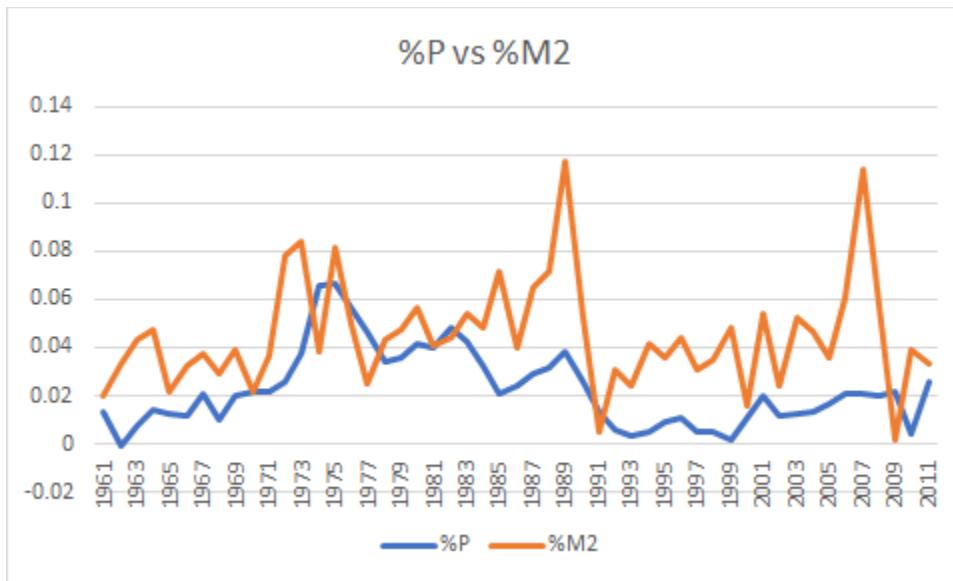
- $\%P = \log P(t)/P(t-1)$
- $\%M1 = \log M1(t)/M1(t-1)$

- $\%M2 = \log M2(t)/M2(t-1)$

According to Friedman, $\%P$ is explained solely by $\%M1$ or $\%M2$. The graphs of these quantities below show clearly that Friedman is wrong



Here the $\%M1$ (growth rate of M1) behaves very erratically compared to prices. The two series have completely different behaviors. It seems clear that knowing $\%M1$ will not help us very much in understanding how $\%P$ behaves. A similar pattern holds for $\%P$ versus $\%M2$ – there is no relationship between the two series.



Because standard statistical and econometric analysis is so complicated, students do not realize that “There is no MAGIC”. If the series LOOK unrelated, they ARE!

Critical Understanding: The GRAPH provides ALL the information that EXISTS. It is a COMPLETE picture of the data. THERE IS NO MORE INFORMATION AVAILABLE. A

Mistaken Concept is created in the minds of students: FANCY TECHNIQUES can allow us to get MORE out of the data. Thus, if we do Limited Information Maximum or Generalized Method of Moments, we might get some MORE information. Even if the picture reveals no relationship, we might FIND OUT that there is strong relationship by applying some complicated techniques.

It is important to understand that ALL FANCY techniques are based on ADDING INVALID COMPLEX assumptions. The ADDITIONAL inferences come from these assumptions and NOT from the data. For example, a standard ARDL (Autoregressive Distributed Lag) analysis will lead us to the conclusion that Real Money Demand is explained by Real GNP, after including sufficient lags of both variables. This is WRONG because it is based on the ASSUMPTIONS of regression model, which are guaranteed to be wrong here. This is not realized by most, because many of the assumptions of the regression can never be tested and proven wrong conclusively.

Concluding Remarks. Pictures of the data provide us with complete information about the data. However, there is an art to drawing pictures. Using the right scale is just the beginning. There are many fancy techniques, now known as “Data Visualization” – these techniques will be an important part of future statistics and econometrics. Computers enable these techniques which were impossible at the time the subject was invented in the early 20th Century. The standard methodology for statistics and econometrics substituted ASSUMPTIONS for analysis, because that was the ONLY path available at that time. Given the difficulty of doing calculations and making graphs, convenient assumptions made analysis unnecessary. This is why a radical change in methods of analysis, and in ways of thinking about data, is necessary, now that extremely difficult computations and extremely complex graphics, can be created with a click.

5: Eugenics and the Birth of Statistics

BLURB – Provide a brief description of overall goals and contents of this chapter

5A Malthusian Approach to Poverty

Preliminary Remarks:

1. I no longer believe in the myth of Objective Knowledge, which I was trained to believe in, during the course of my Western education. When knowledge is objective, then it is the same for everyone, and does not have any relationship to personal experiences and subjectivity.
2. Secular Modernity is a powerful World religion known to mankind. It has defined the goals of life, and ways of living as individuals, communities and nations, for all of us. Because it defines itself as OBJECTIVE, it has deceived us into accepting hidden normative frameworks.

3. It claims objectivity, neutrality and rationality for itself, and a position of privilege in arbitrating disputes among others. Anyone who opposes it is automatically being subjective, emotional, and irrational. Thus it is important to SEE THROUGH this pretense of objectivity of secular modernity.
4. My own position is based on an Islamic Perspective. ALL analysis must HAVE a perspective, and no neutral, objective perspectives exist. Conventional Statistics is really Statistics from a Secular, Modern Perspective (one religion). Here I am offering an alternative perspective

The methodology being used here is that of Michel Foucault, called the Archaeology of Knowledge. We were taught the BINARY theory of truth. Statements are TRUE or FALSE – Objective Knowledge. INSTEAD – look at the EMERGENCE of ideas – IDEAS are the MOST POWERFUL tools in the arsenal of mankind. How these ideas were used to shape history, and how they evolved and changed as a consequence of historical forces. This methodology leads to unique insights not available by any other method. In this sequence of lessons we dig into the foundations of the ideas which led to the creation of modern statistics.

Statistics is presented as an objective and ethically neutral body of tools and techniques. However, it was developed for a very clear, evil, purpose. There are polar Views on Transmission of Knowledge:

- Look at the CHARACTER & Purpose of the transmitters of knowledge
- Look at the CONTENTS – the body of knowledge transmitted

The fundamental underlying question is: “Can the THOUGHTS be separated from the THINKER?”. The balanced position is that BOTH are necessary. We need to Look at the BODY of KNOWLEDGE being transmitted. Also look at the CHARACTER and the PURPOSE of the transmitter/creator of knowledge. In contrast, the Western Intellectual Tradition says: “Look at the SUBJECT MATTER ONLY”. As opposed to this, a popular strand of the EASTERN tradition says: “look at the AUTHORITY of the TRANSMITTER only.” In the West, the MEANING of the word “Probability” changed to reflect this transition – Initially, the word referred to the authority of the transmitter. Later the word was used to mean the weight of the evidence for the matter (see Ian Hacking: The Emergence of Probability).

The Western intellectual tradition defines knowledge as being purely objective, and excludes subjective and personal experiences from the realm of knowledge. This leads to a clear NO answer to the fundamental questions: “Do INTENTIONS of producers of knowledge matter?”, and “Does the CHARACTER of producers of knowledge matter?”. This is in dramatic contrast with Islamic views, according to which ‘Value of Actions Depends on Intentions’. It is easy to see that the nature of scientific knowledge produced depends on the intentions and purpose for which the knowledge is being produced. This topic is now known as the sociology of knowledge, and many simple examples can be given as a proof:

The strong drive for profits led to multi-million dollars for high yield varieties of wheat, but genetically modified to have terminating seeds. This is to enable corporations to sell the same

seed year after year. If the intentions had been to feed the hungry, different types of technologies and seeds would have developed.

- “Orphan drugs” refers to drugs which provide cures to afflictions affecting masses of the poor, who are unable to pay enough to generate a profit on the production of the drug.
- While there is little work on these, massive efforts are being made on designer-drugs, personalized and individually tailored to the genetic structures of billionaires.
- The search for power has shaped the development of war technology, the developments of bombs, missiles and so much else. A humanistic bent would have led to developments of science in different directions.

With this as preliminary, we consider the personalities and intentions of some of the major founding fathers of modern statistics: Sir Francis Galton, Sir Karl Pearson, Sir Ronald Fisher. Today it has been discredited and forgotten, but Eugenics was EMINENT and RESPECTABLE field of knowledge, taught at universities, and with prominent and influential supporters in the early 20th Century. All three of these founding fathers were big names in Eugenics, and statistics was developed by them as a tool to support their Eugenicist views. Eugenics asserts the Racial Superiority of Whites and the inferiority of the other races. Even more, it asserts that the elite classes are genetically superior to the commoners. According to Eugenics, the only Path to Progress lies in the Extermination of Inferior Races (negative Eugenics), and Increasing Growth of the Superior Race (positive Eugenics). Another way to deal with the inferior races was “specialization” – give them roles to fulfill which would fit their limited genetically determined capabilities (for example, making slaves out of the Blacks to do menial chores not demanding intelligence). A brief sketch of the background which led to the emergence of Eugenics as a prominent field of “science” is given below

A convenient point to start is the question of “Why is their poverty?” and “How can we reduce it?”. It will surprise the reader to learn that these are NEW questions. Even though poverty is an age-old phenomena, the idea that this is a social problem which can be, and should be, remedied, is new. For details see [An Islamic Approach to Inequality and Poverty](#). In pre-capitalist world, poverty was not seen as a social problem. Social responsibility, and the idea of a society as one body, where we must take care of each other, was sufficient to deal with problems of poverty. The emphasis on charity in Islamic teachings led to extraordinarily high levels of spending on the poor, especially via the WAQF, which created endowments to deal with different social problems. But similar patterns of charitable institutions for taking care of the poor are seen in all pre-modern societies.

The driver of major change in Europe was the Industrial Revolution which started in England in the 18th Century. The complex circumstances which created this change, and its effects on the transformation of economic, political, and social institutions, are described in [The Great Transformation](#) by Karl Polanyi (and many other books). Of relevance to our current discussion is the fact that industry requires a Labor Force – lives for rent and sale for money. In a capitalist society (like our society today), education is meant to CREATE mindsets suitable for labor force. This means training students to make the goal of life the pursuit of pleasure, power,

and wealth. For this purpose, students are trained to pursue CAREER over other concerns like family, society, spiritual growth, or excellence in any human dimension. For more discussion of this fundamental problem with modern education, see [Learn Who You Are!](#). Social change resulted from the chosen solution to the fundamental problem created by the Industrial Revolution: “**How to create a labor market?**”

The solution created by Europeans was deeply racist. We must believe in TWO classes of people: REAL human beings are capable of enjoying finer things of life. LOW level humans may LOOK like us, but they do not have rich inner lives – they cannot think and feel as deeply. They are closer to animals than to humans. Eugenics is based on the idea that the aristocratic elites have superior genes to the common masses. A little more historical detail is useful in understanding the emergence of these ideas.

A Tale of Two Cities by Charles Dickens opens with a scene where the carriage of a French aristocrat speeding through the crowded streets of Paris crushes a poor child. The aristocrat tosses a few coins to the bereaved mother, and continues on his way without further concern. The extreme inequality, and oppression of the poor by the elites led to the French Revolution, which changed the course of European history. As a consequence of the revolution, there was considerable debate about policies to help the poor, with a view to preventing a similar revolution in England. At this crucial juncture, the ideas of Malthus regarding the causes of, and solution to, the problem of poverty, had a dramatic impact on the policy debate. Humanitarian and compassionate solutions were replaced by cruel and harsh measures to punish the poor for their poverty. Malthus argued that poverty was due to a poor genetic endowment, and was inherited. The problem of poverty arose from the fact that the poor BREED faster than the rich. Thus, poverty is inherited, and being kind to poor, providing social services, is counterproductive. This will only increase the rate of growth of the population of the poor. INSTEAD – we should sterilize the poor, keep them in crowded conditions, encourage spread of disease among them, to keep their numbers low.

This theme became linked with emerging theories of evolution and Mendelian genetics. How much of our makeup comes from inheritance, and how much is due to the environment and education we receive? This is often called the Nature Versus Nurture DEBATE. The dominant and widely accepted point of view leaned heavily in favor of Nature (heredity) having an overwhelming effect. This means that superior races will remain superior, and it is not possible to educate the inferior races to bring them up to the standards of the white people. The inferior people can either be exterminated or enslaved. The Malthusian approach to poverty was strongly based on the “nature” point of view. That is, poverty was due to bad genes which led to poor character and intelligence, and there was nothing we could do to change this, in the form of education or other interventions.

The wrong theories of Malthus led to a dramatically wrong approach to poverty. Malthusian theory suggested that providing support to the poor would only increase poverty, since that would allow them to breed faster. Thus social support for the poor was made deliberately humiliating and degrading, to discourage all but the extremely needy to resort to the poor houses. These theories and policies stand in stark contrast to Islamic teachings which urge

us to provide support to the poor without humiliating them in the process. Furthermore, Islam teaches us that every human life – whether poor or rich, black or white, Arab or other – is equally precious. Indeed each human life counts as heavily as all of humankind. There were similar humanitarian streams of thought in the European Christian heritage, but Malthusian views came to dominate policy.

Even though nearly all of the predictions of Malthus turned out to be wrong, his theories had tremendous impacts on thinking about population. As we will see in the next portion of this lecture, the founding fathers of modern statistics invented the subject, tools, and techniques, in an attempt to prove the theories of Malthus. Our main goal is to show that tools developed have been influenced by the underlying agenda, and are not neutral and objective. Malthus created his theories without a shred of empirical support, purely from his imagination. Since then, the theories of Malthus have been decisively proven wrong by the empirical evidence. The article, [Malthus: the False Prophet](#), from the Economist, documents some of the major errors made by Malthus.

1. Malthus argues that the population would increase geometrically. However, over the 20th Century, a Demographic Transition was observed, when increasing prosperity, and increasing likelihood of survival of children led to a reduction in birth rates, and stable population sizes.
2. Malthus argued that food supplies would increase linearly, leading to shortages. However, continuing sequences of technological advances in agriculture have led to increasing food supplies per capita on a global basis.
3. There were many other false predictions, based on the first two. For example, he argued in 1798 that Britain population would quadruple in 50 years to 28 million, but food supplies would only be sufficient for 21 million, leading to a crisis. But nothing remotely resembling this happened.

Despite numerous fallacies and failed forecasts, the ideas of Malthus continue to be exceedingly popular. WHY? The simple answer is that these theories are ALIGNED with class interests of the RICH. The POOR are to blame for their poverty. Furthermore, the rich have no responsibility to help the poor, because helping them only increases their breeding rate, leading to increased poverty, as well as increasing the stock of bad genes in human population. These deeply mistaken ideas, strongly in conflict with Islamic teachings, have had a deep and disastrous impact on human history, adding to misery of millions. They continue to guide thinking and policy of an influential minority of economists and politicians. In the next portion of this lecture, we look at the development of statistics as a tool of Eugenics, a field of study built on the foundational ideas of Malthus and Darwin.

5B: Sir Francis Galton: Eugenicist Founder of Statistics

The Islamic tradition asks us to look at both the nature of the knowledge, as well as the character and intentions of the transmitters of knowledge. In this lecture, we will look at Sir Francis Galton, the Founder of Eugenics. The following quote from his student and admirer Karl Pearson (1930, p. 220) explains Eugenics:

“The garden of humanity is very full of weeds, nurture will never transform them into flowers; the eugenist calls upon the rulers of mankind to see that there shall be space in the garden, freed of weeds, for individuals and races of finer growth to develop with the full bloom possible to their species.”

Looking through the metaphor of flowers (Europeans) and weeds (others), Eugenics call for the EXTERMINATION or STERILIZATION of inferior races, as well as inferior specimens among the Aryan (White) race. This reflects WIDELY HELD views among Europeans. NOTE the conflict with WISDOM of Quran: All Human Beings are Brothers and Sisters, Sons & Daughters of Adam and Eve. Furthermore, ALL human lives are infinitely precious – each life counts as heavily as the entire humankind.

The value of actions depends on intentions. An essential part of the Islamic approach to the acquisition of knowledge involves making the intention of serving the creation of God, out of the love of the Creator. Evil intentions lead to bad outcomes. In particular, some evil intentions are mentioned in the Hadeeth as follows: He who seeks knowledge to argue with Experts, to dispute with the ignorant, or to attract attention (seek popularity), all his deeds will be in vain.

We can document the intentions of Galton were exactly the ones which are forbidden for Muslims. He sought popularity and fame, to enter into disputes with the experts to impress the ignorant. For documentation, see [Becoming a Darwinian: The Micro-Politics of Sir Francis Galton's Scientific Career 1859–65](#). These intentions had a strong effect on the quality of knowledge produced – creating wrong paths of research and wrong ways of thinking, which continue to influence production of knowledge in the field of statistics in harmful directions. We discuss this in somewhat greater detail below.

Sir Francis Galton was the founder of Eugenics, while both Sir Karl Pearson, and Sir Ronald Fisher were prominent members of the deeply Racist movement of Eugenics. Their research was meant to create a Scientific basis for Racism. This had two aspects.

1. *Positive Eugenics*: Use of breeding to create a superior stock of human beings.
2. *Negative Eugenics*: Extermination or Specialization of Inferior Races. Specialization refers to assigning some subhuman and subservient role to a race – such as assigning blacks to be slaves of the Master Races.

The consequences of these ideas were horrific, leading to sterilization of thousands, culminating in the brutal killing of millions of Jews by Hitler. Because of these shameful consequences, even the name of the field has been erased from history. For one look at the evil consequences of the idea (which is current and dominant) that man is just another species of animal, see [The Darwin Effect by Jerry Bergmann](#). The idea that mass killing of inferior peoples

is actually necessary to reach an advanced state of civilization has been expressed by Karl Pearson, successor of Galton, and one of the eminent founders of statistics, as follows:

"History shows me one way, and one way only, in which a high state of civilization has been produced, namely, the struggle of race with race, and the survival of the physically and mentally fitter race. If you want to know whether the lower races of man can evolve a higher type, I fear the only course is to leave them to fight it out among themselves, and even then the struggle for existence between individual and individual, between tribe and tribe, may not be supported by that physical selection due to a particular climate on which probably so much of the Aryan's success depended." (Karl Pearson, 1901, pp. 19-20)

What were the statistical tools created to support the cause of Eugenics, or mass extermination of "inferior" people? Galton invented "correlation". Why was Galton trying to measure relationship between heights of fathers and sons? He wanted to prove that heredity was all important in determining the heights of sons. But how does height relate to personality and intelligence? We must understand the strong hold of materialism on 19th Century minds. Everything was matter and was observable. Things which could not be observed did not exist. Weighing a body before and after death, scientists concluded that the soul did not exist, because the weight did not change. Thoughts were considered to be fluid secretions of the brain. Phrenology was in vogue – this pseudo-scientific field took measurements of the skull to determine personality. So if physical characteristics were hereditary, it would be sufficient to establish that personality, intelligence, character, etc. were also hereditary. This was the contribution of Galton – whereas Darwin's followers had only considered physical characteristics, Galton argued that intelligence and personality were also hereditary, and subject to evolutionary pressures. This is also called "social Darwinism", where it was argued that human societies could evolve to become better, by the ruthless survival-of-the-fittest mechanism, which eliminate the weakest members, to strengthen the race.

We look briefly at some technical details of Galton's ideas about correlation and heredity. For any characteristic – like IQ, height, strength, etc. – we can subdivide the population into three groups: High, Medium, and Low. The question of correlation between sons and fathers was subject of intensive research by Galton and his followers. The following Diagram shows PERFECT correlation, the type of result that they were hoping to find:

This diagram shows 100% Correlation: High IQ fathers always have High IQ sons. Similarly, the characteristics of the parents are transmitted to the children perfectly, for mid-level and low-level IQ as well. If the effect of heredity is strong, this justifies BREEDING human beings, like dogs and horses, to create a stock of SUPERMEN. Select High IQ people, ensure that they mate with each other, and sterilize or exterminate the rest of the population. This was one of the major GOALS of Eugenics.

To achieve clarity in understanding any concept, it is always useful to look at the opposite alternative: NO inheritance. This can be represented in the following diagram:

Here regardless of any parentage, the children are equally divided among the three categories. There is no effect of heredity. Even when there is 100% correlation, we cannot

establish that this is due to heredity because smarter and wealthier parents are able to provide a better educational environment for their children. The Eugenics arguments becomes much weaker if the effects of heredity are weak. If children of any type of parents can achieve any level of intelligence via training and education, then breeding for intelligence becomes impossible. Eugenicists were stubbornly opposed to this idea, and resisted any interpretation of empirical evidence against heredity and in favor of environment.

The teachings of Islam, and the example of our Prophet Mohammad SAW show us that every human being is valuable, because everyone has a soul, and the capability to know God. With appropriate training, everyone can achieve high levels of spirituality. No one is born with these traits, and it requires struggle against the desires of the Nafs to advance from primitive spiritual stages to the higher ones. Anyone with some levelof spiritual development would have known that the idea of breeding human beings was patently ridiculous. The higher stages of human development can only be achieved by struggle, and all human beings are born with the capability of carrying out this struggle. The research program of the Eugenicists was possible only because of the low level of spiritual development of the founders of statistics. Human beings who make their desires their God reduce themselves to the lowest of the low, and make themselves similar to animals. Only then it becomes possible to consider breeding humans like animals and killing them like animals. This was the effect of widespread materialism in Europe.

The weaknesses of the tools in statistics arose because the Eugenicists were out to prove something which was not true. Only bad tools could accomplish this goal – if statistics had been constructed on the correct foundations, they would have been unable to prove their favorite theses. We provide some further description of these twisted tools developed to achieve racist goals.

We have shown the graphical version of perfect versus zero correlation in Inherited Characteristics. An alternative way of showing the same quantitatively is via Markov Transition Probabilities. Here the first table below shows the case of perfect correlation:

| F⁻ S® | High | Medium | Low |
|-------------------------|-------------|---------------|------------|
| High | 100% | 0 | 0 |
| Medium | 0 | 100% | 0 |
| Low | 0 | 0 | 100% |

The rows show the characteristic of the fathers, while the columns are the sons. The transition probabilities show that 100% of High IQ fathers have High IQ sons, and similarly for Medium and Low IQ. There is perfect correlation between fathers and sons. The next table shows the case of ZERO correlation:

| F⁻ S® | High | Medium | Low |
|-------------------------|-------------|---------------|------------|
| | | | |

| | | | |
|--------|--------|--------|--------|
| High | 33.33% | 33.33% | 33.33% |
| Medium | 33.33% | 33.33% | 33.33% |
| Low | 33.33% | 33.33% | 33.33% |

In this case, regardless of the type of father, all sons are equally divided into the three possibilities of High, Medium, and Low IQ. Empirical evidence regarding heights shows that there is roughly a 50% correlation, which corresponds to a 50-50 mix of the two polar case of perfect and zero correlation. This can be displayed in the following table:

| F ⁻ S® | High | Medium | Low |
|-------------------|--------|--------|--------|
| High | 66.67% | 16.67% | 16.67% |
| Medium | 16.67% | 66.67% | 16.67% |
| Low | 16.67% | 16.67% | 66.67% |

Here the majority 2/3 of children fall in the same category as the father, while the rest are equally divided among the other two categories – this equal division may be replaced by other assumptions, and does not matter for what follows. What we would like to show is that, in all cases OTHER than perfect correlation, we will see the phenomenon of “Regression Towards the Mean”. Note that if High IQ parents do not have 100% High IQ children, then they will necessarily have children of LOWER intelligence. That is the ADVANCED intelligence will be REDUCED towards the average, or medium, intelligence. Similarly, in the LOW category, if all children are not LOW IQ than they will have HIGHER IQ and therefore move UP towards the mean. That is, there is a tendency of both extremes to move towards the normal. This “Regression towards Mean” causes problems for Eugenics. We cannot rely on High IQ parents to produce High IQ children based on heredity. ALSO, there is no reason to sterilize or exterminate Low IQ parents, because their children may move UP the IQ scale. Nonetheless, deprived of the light of the message of the Quran about human equality and brotherhood, Eugenicists continued to mis-interpret empirical evidence, and advocate selective breeding of superiors and extermination of inferiors. The horrors of the Holocaust, where millions of innocent men, women, and children, were burnt alive, eventually discredited the cause of Eugenics.

Conclusions

Ideas are far more powerful than atom bombs – after all, it was ideas which led to the creation of atom bombs. The idea that man is an ANIMAL (Darwinism) has led to a lot of damage. By denying the existence of spirituality, It has made spiritual progress impossible. This idea has penetrated minds of Muslims through Western education, which is purely materialistic.

The potential for excellence in men is unlimited, and this can only be achieved through spiritual training – This CANNOT be inherited. Failure to recognize this potential reduces us to animals. This conception of man as an animal is at the heart of the Social Sciences developed in the West over the past few centuries. Because of this fundamental flaw at the roots, all productions of knowledge in this intellectual tradition are tainted. Statistics forms one part of the tools used to try and validate the racist ideas of the originators of the subject. Although it has moved far from these roots, and has many valuable accomplishments and ideas to its credit, the current shape of the subject continues to be affected by its origins, as we will show in later lectures.

5C: Fisher's Failings and the Foundations of Statistics

Fisher was a prominent Eugenicist, and he had six children in accordance with his belief that the path to improvement of the human race involved increasing the propagation of superior specimens of humanity. A central question for us is: "Is modern statistics FREE of its Eugenicist origins?". The minority position is NO. This position is described and well defended by Donald Mackenzie in his book: "Statistics in Britain, 1865 to 1930: The Social Construction of Scientific Knowledge". He writes that "Connections between eugenics and statistics can be seen both at the organisation level and at the detailed level of the mathematics of regression and association discussed in chapters 3 and 7. Without eugenics, statistical theory would not have developed in the way it did in Britain - and indeed might not have developed at all, at least till much later." In brief, Eugenics shaped the tools and techniques developed in statistics. However, the Dominant View is that Modern Statistics is FREE of its racist origins. This view is ably defended by Louçã, Francisco in his article on "Emancipation Through Interaction—How Eugenics and Statistics Converged and Diverged." Journal of the History of Biology 42.4 (2009): 649-684. He argues in favor of the Consensus View: There is no doubt that origins of statistics are due to Eugenics project, but it has now broken free of these dark origins.

In this part of the lecture, we look at the personality of Fisher, and assess how it shaped the foundations of statistics. It is acknowledged by all that Fisher was cantankerous, proud & obstinate. He would never admit to mistake, and was stubborn in defending his position, even against facts. He was also vengeful: To oppose Fisher was to turn him into a permanent enemy. In many battles, Fisher took the wrong side. HOWEVER, he won most of his battles because of his brilliance, to the detriment of truth. The impact of Fisher's victories has permanently scarred statistics, and continue to guide the field in the wrong directions. This lecture is about SOME (not all) of his fundamental mistakes.

Perhaps the most basic, and also the most confusing, was the battle between Fisher and Pearson regarding the testing of Statistical Hypothesis. This is confusing because today both of the two conflicting positions are taught to students of statistics simultaneously. Even though the conflict was never resolved, it is now ignored and glossed over, buried under the carpet. The Fundamental Question is "WHAT is a hypothesis about the data?". According to Fisher, a hypothesis treats data as a random sample from a hypothetical infinite population which can be described by a FEW parameters. WHERE does this ASSUMPTION come from? It comes from the NEED to reduce a large amount of data to a FEW numbers which can be studied. This

reduction is needed because of our LIMITED mental capabilities – we cannot handle/understand large data sets. Fisher wrote that: “*In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.*” The parameteric mathematical model for treating the data as a random sample from a hypothetical infinite population allows us to reduce that data, making inference possible. The hypothetical infinite population does not have any counterpart in reality.

What is to prevent the statistician from making completely ridiculous assumptions, since the model comes purely from the imagination, and purely for mathematical convenience? For this purpose, Fisher proposed the use of p-values. If the data is extremely unlikely under the null hypothesis, this casts doubt on the validity of the proposed model for the data. The p-value tests for GROSS CONFLICT between data and the assumed model. One can never learn whether or not the model is true, because there is nothing real which maps into the assumed hypothetical infinite population which follows the theoretical distribution being assumed. To Fisher, the mathematical model is a device to enable the reduction of the data, and not an true description of reality.

In a classical example of mistaking the map for the territory, the Neyman-Pearson theory of hypothesis testing takes the Fisherian model as the TRUTH. The Null hypothesis is ONE of the parametric configurations. The Alternative hypothesis is SOME OTHER parametric configuration. The Neyman-Pearson theory now allows us to calculate the exact most powerful test – under the assumptions that the parametric models COVER the truth. The possibility of a TYPE III errors – that is, none of the assumed parametric models is valid – is ruled out by assumption, and never taken into consideration. BUT the assumption of a parametric model to describe the data is arbitrary. The imaginary infinite population following a theoretical distribution has been made up just for mathematical convenience!

In the course of the bitter personal conflict which ensued, the real issues, related to the common weakness of both approaches were ignored and suppressed. Instead, Fisher’s promotion of his methods led to dramatic misuse & abuse of the Fisherian p-values. The P-values MEANT to assess gross conflict and serve as a rough check on the modelling process. Instead, these were turned into a REQUIREMENT for valid statistical results. The hugely popular philosophy of science developed by Karl Popper was very useful in elevating the importance of the p-value: we can never PROVE a scientific hypothesis, but we can disprove them. A significant p-value disproves a null hypothesis creating a scientific fact. Insignificant p-values mean nothing. This led a fundamentally flawed statistical methodology currently being taught and used all over the world. The problem is that there are huge numbers of hypothesis which are NOT in gross conflict with the data. By careful choice of parametric models, we can ensure that our desired null hypothesis does not conflict with the data. The Neyman-Pearson theory can ADD to this

illusion of the validity of imaginary hypothesis, if we find alternatives which are even more implausible than our favored null hypothesis.

Fisher Versus Gosset. The p-value invented by Gosset measures statistical significance, which is very different from practical significance. Gosset warned against confusing the two from the beginning. Unfortunately, because it was a tool in Fisher's war against Neyman-Pearson, Fisher pushed it to the hilt. This led to a fundamental misunderstanding of the role and importance of p-values in statistical research which persists to this day. The damage inflicted by these misguided statistical procedures has been documented by Stephen T. Ziliak and Deirdre N. McCloskey in **The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives**

Perhaps of even greater fundamental importance was the battle between Fisher and Sewall Wright. Sewall Wright invented path analysis – a method for assessing CAUSAL effects. If this method had been understood and adopted, modern statistics would be entirely different. Unfortunately, Sewall Wright had a fight with Fisher on some obscure genetics controversy related to EUGENICS. As a result, Fisher's ignored, neglected, and criticized, all contributions, and attempts at developing a theory of causality. To be fair, this was not entirely Fisher's fault. Theories of knowledge in vogue, based on logical positivism, suggested that unobservables cannot be part of scientific theories. This led to difficulties in understanding causality, because it is never directly observable, and is always based on understanding of unobservable real-world mechanisms. Over the past few decades, there have been revolutionary advances in understanding of causality, made by Judea Pearl and his students, which build on causal path analysis similar to the methods of Sewall Wright. Unfortunately, statisticians and econometricians have mostly failed to learn from these methods, because they go against decades of indoctrination against such methods.

Failure to understand causality continues to be a serious problem for statistics. One of the most dramatic illustrations was the controversy about Cigarettes and Cancer in the middle of the 20th Century. For more details about this controversy, see Pearl & Mackenzie: **The Book of Why** (Chapter 5) and also [Walter Bodmer: RA Fisher, statistician and geneticist extraordinary: a personal view](#). A friendly relationship turned into enmity when Bradford Hill and Richard Doll published an extensive empirical study documenting the effect of smoking on cancer. This conflicted with Fisher's views that correlations cannot prove causation, and also ideology of libertarianism. These convictions led Fisher to deny empirical evidence regarding the link between smoking and cancer long after it had become overwhelming. Because of his enormous prestige, his opinions delayed recognition of the link, and the necessary policy response, which led to decline in smoking and deaths from cancer.

What lessons can be learned from this personal history of the founder of modern statistics? Islam teaches us a lot about the search for knowledge. See Principles of Islamic Education for a detailed discussion. Here we briefly discuss some of the required attitudes for Seekers of Truth. We must learn to valuing knowledge as the most precious treasure of God, seeking it with passion, energy, and utmost effort. This was one of the keys to how Islamic teachings made world leaders out of ignorant and backwards Bedouin. We must also understand

that knowledge, or insight, is a GIFT of God. We must learn to take small steps, and be grateful for small advances in understanding. Knowledge is like a castle constructed brick-by-brick from small elements. We must acquire patience for the long haul, instead of expecting quick results. The knowledge we acquire does not come from our personal capabilities; it is a gift of God. We cannot take pride in discoveries because they are not due to my genius. The pride of Qaroon is condemned – that my wealth is due to my own wisdom and capabilities, and therefore I do not recognize the rights of others. We must learn humility & gratitude: I have been given knowledge beyond what I deserve, and beyond my capabilities. Furthermore, because of our limited capabilities, we can often make mistakes, and fail to recognize the truth, and confuse it with falsehood. Thus, we must ask Allah to show us the difference:

اللهم ارنا الحق حقاً وارزقنا اتباعه وارنا الباطل باطلاً وارزقنا اجتنابه



Human knowledge is a social construct. That is, we create knowledge collective, and validity is created by consensus. This makes it tremendously important to have the right rules for discussion and argumentation. Islam has a well developed collection of rules which govern the Etiquette of Discourse. Unfortunately, these are no longer studied and taught, much less practiced. Among the rules is the prohibition of *Debate*, with a view to prove ME right and YOU wrong, and WIN arguments. This just feeds the ego and creates pride. Instead, arguments are SUPPOSED to be a cooperative search for truth. To explain this point, a teacher told his student never to debate. The student was surprised and said – “But we have seen you debate?”. The teacher explained that in the earlier generations, people entering into an argument would make dua “O Allah, make the Haq clear to us, and let the truth and clarity come from the OTHER person.” In my generation, people would ask for the Haq to be made clear, but also ask that the truth should come from MYSELF. In your generation, people have stopped caring about the truth, and only seek to win arguments!

Conclusions

To summarize this discussion, we conclude that modern conventional statistics is based on fundamentally flawed methods. Fisher created a method for reducing data based on an imaginary infinite population. Because this is imaginary, there is no possibility of assessing the truth of such a hypothesis. However, if it is strongly in conflict with apparent characteristics of

the data, the p-value can be used to reject such a null hypothesis. A vast class of null hypothesis will fail to be in manifest conflict with the data, and failure to reject is the closest we can get to truth in Fisherian methodology. This was useful for the cause of Eugenics, since it can allow us to prove MANIFESTLY false null hypotheses. The methodology for statistics started out in the wrong place, partly due to the Eugenic roots, but also due to computational limitations and to the personality of Fisher. However, the persistence of these flaws across time was due to failures in the ethics of dialog – the rules for social construction of knowledge. The followers of Fisher failed to consider the origins of the methods, and imitated his methods without understanding them. In particular, they failed to understand the radical revisions of methodology made possible and necessary by the amazing advances in computing capabilities. This is why it has become necessary to re-build statistics on new foundations, abandoning the Fisherian methodology. We will discuss some of reforms required in later portions of this lecture.

5D: Real Statistics: Alternative to Fisher's Approach

In previous lectures, we have explored some of the reasons why foundations of modern statistics constructed by Sir Ronald Fisher are deeply flawed. In this lecture we explain the basics of our alternative approach to the subject.

This lecture will explain how we can re-build Statistics on new foundations. To do this, we will first explain the foundations of conventional statistics – which may be called “nominalist” or Fisherian statistics. Then we will explain the alternative approach we propose, naming it REAL statistics. Our goal in this lecture is to provide clarity on the differences between the two approaches.

The Fisherian approach is based on fancy mathematical models, which are purely IMAGINARY – That is, the models come from the imagination of the statistician, and have no corresponding object in reality against which they can be verified. A Fisherian MODEL for data ALWAYS involves treating SOMETHING as a perfectly random sample from a hypothetical infinite population. However, there is flexibility in what that “something” may be – it is this flexibility that is deadly, allowing us to prove anything we like. The flexibility was not originally part of Fisher’s approach. He proposed to model the data directly. Later workers “generalized” his approach to make it applicable to a wide variety of data sets and situations. This generalization was dangerous because it makes unverifiable assumptions about unobservable entities and uses this as the basic engine of inference. In contrast, the Fisherian approach makes assumptions directly about the data, and hence is easier to assess and understand, although equally difficult to prove or disprove.

The typical use of this imaginary methodology involves breaking the data into two components: DATA = LAW + ERROR. The LAW captures a flexible class of models which you believe to be true. This flexibility makes the ERROR unobservable, because it shifts as you try out different potential laws. This gives you a HUGE potential for constructing ANY LAW you

like to explain the data – what is unexplained by the law is AUTOMATICALLY part of the ERROR.

We illustrate how this methodology allows us to prove anything at all: Take any data, and decompose it as DATA = Desired Law + Error. This is always valid By DEFINING Error := DATA – Desired Law. Now make STOCHASTIC assumptions about Error in rough conformity with errors obtained at your desired law. Current methodology allows us to make almost any assumptions we like about the error. The beauty of the stochastic assumptions is that a wide range of numbers satisfy them. If we say that the errors follow some common distribution (that is, they are random draws from a hypothetical infinite population) it is very hard to assess whether or not this is true. This difficulty is increased because a flexible range of laws make it difficult to pinpoint the actual errors, to check the stochastic assumptions. Furthermore conventional methodology generally does not bother to even try to test assumptions on errors, making it even easier to prove any model conforms to the data.

The key illusion created by conventional statistical methods is based on a misunderstanding of the nature of statistical models. ALL statistical inference is based on the IMAGINARY stochastic model regarding errors. HOWEVER, textbooks create the widespread belief that inference comes from the DATA! This is what permits us to “LIE with statistics”. Making complex assumptions about errors allows us to achieve any kind of inference, and attribute this to the data. Then we can browbeat people by telling them that we have made a deep analysis of the data, and the truths we have uncovered cannot be accessed by ordinary people not trained in the mysteries of sufficient statistics. In fact, the inferences come from unverifiable assumptions about unobservable errors.

In opposition to this, we propose an alternative, which we will call REAL Statistics. At the heart of this approach is the idea that the data provides us with CLUES about underlying realities. The goal of inference is NOT related to DATA itself. Rather, the GOAL is to use the data to UNDERSTAND the real-world processes which generated the data. This NECESSARILY involves going beyond the data. Conventional statistics treats only the data, and Fisher explained that the goal of statistics is to reduce large and complex data sets to a few numbers which adequately summarize the data and can be understood. Today, because of advanced computational capabilities, we are able to directly handle large data sets, and can move beyond this idea of statistics as being the reduction of data sets.

Our approach radically changes the task of the teacher of statistics, requiring the creation of new textbooks as well as more training. We must ALWAYS look at the DATA set together with the REAL WORLD PROBLEM under study with the help of the given data set. We can NEVER study DATA sets in isolation, as a collection of numbers. This teachers will have to acquire knowledge and expertise going beyond the numbers to the real world phenomena which generate the numbers.

Another way to understand conventional statistics is to say that it has the following GOAL: find STOCHASTIC patterns in the data. These patterns allow us to treat the data as Random Sample from an IMAGINED population. There is NO WAY to assess validity of

imaginary assumption. The pattern is in the eye of the beholder, and cannot be matched against real structures to see whether it is “true”. The standard methods to assess validity of patterns are goodness of fit, prediction, and control. These are central to conventional methodology, but of peripheral interest in the real methodology. To understand why the search for patterns fail, we consider the failure of this methodology illustrated by the failure of the forecast competition run by International Journal of Forecasting (IJoF). The IJoF ran a competition for many years, where researchers were invited to submit algorithms for finding patterns in data, and using these algorithms to predict the next few data points. IJoF tried these different pattern finding algorithms on a thousands of real world data series to see which one works best. But these competitions did not yield any consistent results. Different types of algorithms would perform differently across series, with unpredictable patterns of performance. This becomes perfectly understandable from the REAL statistics perspective. An algorithm would perform well if and only if the pattern it discovered matched the underlying real world structures which generate the data. These structures differ widely across the data series and so no one algorithm could find them all. It is only after we know the real world context that we can search for the right kind of pattern. Without checking for match to reality, we are just ‘shooting in the dark’ and completely random forecasting results are to be expected. For more details, see [A Realist Approach to Econometrics](#).

We come to the question of “How to do REAL statistics?”. The basic goal is to Look at the BEHAVIOR of the data to get CLUES about the operation of the real world. Note that this step – looking at the data – was NOT POSSIBLE when Fisher created his brilliant methodology (for the time). Given a 1000 points of data, it was a massively laborious task to graph the data, or to create histograms, which provide a picture of the data distribution. Now, we can do this with one click. The ultimate GOAL is to discover CAUSAL EFFECTS, or UNOBSERVABLE OBJECTS, which give rise to the patterns we see in the data. But the first step is to just be able to look at the patterns in the data, without imposing preconceived patterns on them, as required by the Fisherian approach. Descriptive Statistics is about LEARNING to look at the data in a way which leads to LEARNING about the real world. The real world is characterized by unobservable objects and unobservable causes. But before we can learn about these deeper realities, we must learn how to read the surface – the appearance of the data. An early approach to “just looking at the data” was pioneered by Tukey, with the name of Exploratory Data Analysis. EDA was a collection of techniques for looking at the data. However, it was consistent with, and complementary to, the Fisherian approach. The goal was to see if the data patterns would validate a Fisherian model for the data, or whether they would suggest some alternative theoretical models. EDA looks at the data in order to generate a Fisherian hypothesis about the data – NOT a hypothesis about the real world process which generates the data.

The TASK of a DS teacher is much more difficult than that of a conventional statistician. Biometrics is the study of statistics applied to Biological Problems. The teacher must know some biology in addition to statistics. The real world context has dramatic effect on HOW to analyze the numbers. We illustrated this by the study of inflation, where the discussion required understanding WHY inflation matters, and WHY we are trying to measure this. Different

numbers and different techniques become useful according to different uses for these inflation numbers.

Since there is no universal collection of methods valid for all contexts, teaching can be by apprenticeship, via case studies only. Within any real world context, we must learn about the real world to understand the linkages between the real world and the numbers which measure aspects of the real world. We must know the MEANING of the numbers, not just the numbers. This necessarily requires going beyond conventional statistics, which deals only with analysis of numbers. No template for analysis can be given to students. Rather, by teaching how to think about numbers in different real world contexts, we hope the student will learn some ways of thinking which can be applied more generally. This is like the “case study” method now popular in business schools. In this course, we will illustrate this methodology in different contexts.

Concluding Remarks

In this course, we are trying to learn HOW to LOOK at DATA. This is because this is first and introductory course. Learning and analyzing deeper real world objects and causes is very much a part of REAL statistics, but requires advanced methods, suitable for later courses. We note that techniques of “Data Visualization” enabled by computers were far beyond the reach of researchers a few decades ago. Making a histogram, or a graph, of 1000 data points was extremely laborious task. Now it can be done with a click. It is NO LONGER necessary to make convenient simplifying assumptions – as in Fisherian approach to statistics. This leads to a radical conclusion: A HUGE amounts of extremely sophisticated mathematical theory is PURELY IMAGINARY and can be thrown out of the window! , We can temper this radical conclusion by noting that there are certain limited contexts where the Fisherian probability models provide an adequate match, or even an excellent match, to the actual data. In such cases, the original methods would continue to be valid and useful, as supplements to the more general approaches to be studied in Real Statistics.

5E: Contrasts between Fisher's Approach and Real Statistics: The Case of Inflation

In previous portions of this lecture, we have emphasized the need for a new approach, which we call “Real Statistics”. In this lecture, we illustrate the differences between the conventional approach and our new approach using the already studied example of Australian Inflation. In this connection, it is of great importance to understand the following:

The DATA is ALL we have – The STATISTICAL ASSUMPTIONS imposed on the data DO NOT PROVIDE US with additional information. HOWEVER, all statistical inferences we make RELIES HEAVILY on these UNVERIFIABLE (and typically false) ASSUMPTIONS.

First Step of a REAL analysis: LOOK at the DATA with reference to a REAL world issue under examination. In this case, we are interested in the Quantity Theory of Money in general. In particular, we want to examine Milton Friedman’s idea that “Inflation is always and

everywhere a monetary phenomenon, in the sense that it is and can be produced only by a more rapid increase in the quantity of money than in output" The data can tell us whether or not this important hypothesis about the economy, which asserts the neutrality of money, is true.

Graph of Prices (GDP Deflator) and Broad Money. Money has been rescaled to be 100 in 2019, just like the price index series. This data is taken from the WDI data set.

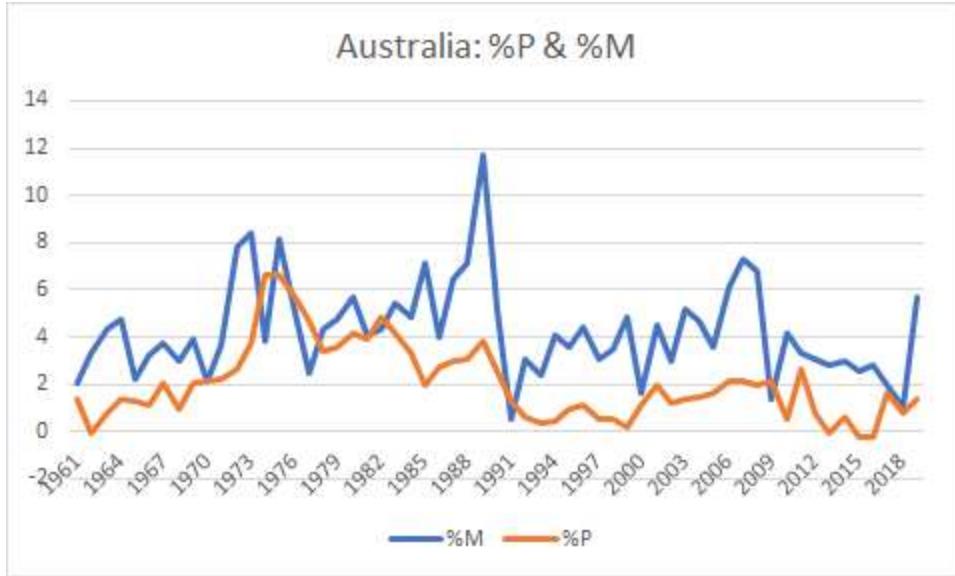


The two graphs clearly show different trends. In the early period from 1972 to 1990, prices are increasing sharply, while money is increasing slowly. Later, Money starts to increase sharply while price curve is flatter, showing a smaller rate of growth. Looking at the graph leads to the IMMEDIATE conclusion: There is no strong direct relationship between money and prices. Note that this conclusion is based on direct examination of data, without any stochastic assumptions required for the Fisherian approach.

Second Step: Look at DATA in WAYS which are relevant to ISSUE of concern! In this case, the QTM tells us the increases in money stock lead to increases in prices. To examine this, we need to look at the Rate of CHANGE in prices, and also the Rate of CHANGE in money. In previous analysis, we came to the conclusion that the best measure of rate of change is the following:

- Define $\%P = \log\{P(t)/P(t-1)\}$
- Define $\%M = \log\{M(t)/M(t-1)\}$

With this definition, the growth rate of money over two years will be sum of the separate growth rate for each of the two years. A Graph of $\%P$ and $\%M$ is given below:



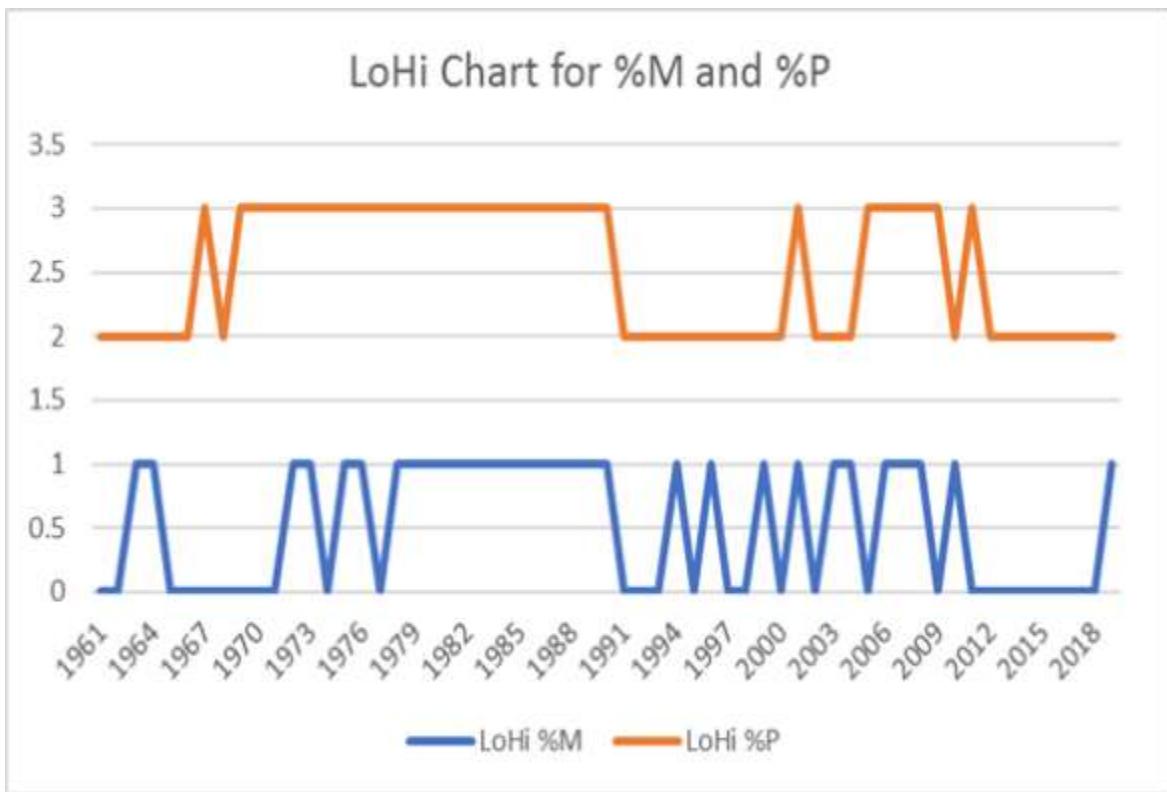
This graph shows a ROUGH correspondence between the two series, but also shows many anomalies. That is, there appear to be sharp increases and decreases in %P which do not correspond to any similar change in %M. It does not seem that %M can explain all the fluctuations in %P, contrary to Friedman's dictum. However, this is just a preliminary impression, and more careful analysis is needed to come to firm conclusions.

Third Step: Find ways to analyze the data SUITED to the question you are asking. The jagged graph above contains too much information, and is not directly suited to telling us about: "How STRONG is the ASSOCIATION (not causation) between %M and %P?" Note that association is symmetric, and causation is uni-directional. Even though we are interested in the causality – does %M cause %P or is it the reverse? – the data cannot tell us about this crucial question. Techniques for studying causality are extremely important, but are not part of conventional, Fisherian statistics. In fact, many famous statisticians are on the record as having denied the relevance, importance, or even meaningfulness of the idea of causality. We will not look at this debate in any depth in this current course, which deals with elementary concepts only. However, causality must be an important part of any "REAL" statistics.

As a first step towards the deeper and more complex concept of causation, we can try to measure contemporaneous association. "Contemporaneous" means that we look at the relation between $\%P(t)$ and $\%M(t)$ for the same year t – we do not look at associations across time. This is always a useful first step. The STANDARD METHOD in use for this purpose is as follows. ASSUME data is jointly normal. Apply the formula for the correlation coefficient; this is the best measure of association for Bivariate Normal Distribution. As with all methods of conventional Fisherian statistics, this method suffers from a serious PROBLEM: it works VERY POORLY if data is not Normal. There is NO REASON to assume data under examination is like a random sample from a hypothetical infinite bivariate normal population. Instead, we develop a direct and intuitive method for evaluating association below.

As a first step, divide %P and %M into HIGH and LOW. We want to know if %P is High when %M is High, and also whether %P is low when %M is Low. The question is: How to divide a series into HIGH and LOW parts? There is a NATURAL and INTUITIVE methodology for doing so. We can SORT the series in ascending order in EXCEL. Find the MIDPOINT. We have an annual series with 59 point of data, so the 30th data point would be in the middle. Series data points BELOW midpoint can be classified as LOW, while data points above midpoint are HIGH. This is natural in the sense that the 29 lowest rates of %P within the 59 points are classified as LOW, and the 30 highest data points within the data set are classified as high.

Using this method of classifying %P and %M into HIGH and LOW values, we can make a chart of Lows & Highs for %M and %P as follows:



Many interesting patterns can be seen from the above graph, which shows the highs and lows for both %P (changes in prices) and %M (changes in broad money). First we note that from 1961 to 1972 rates of money growth (%M) were LOW, except in the two years 1963 and 1964. Corresponding to this period, we find that %P was low in 1961-66, High in 1967, Low in 1968, and then consistently High from 1969 to 1990. There were two decades of high inflation in 70's and 80's was followed by a low inflation period from 1991-2004. This picture immediately leads to many questions:

Why was there an episode of high inflation in 1967? If Friedman's hypothesis is true, than it must have been due to previous high rate of increase in money. Could it be that High %M

in 1963, 1964 led to High %P in 1967? Knowing about the mechanisms of money and prices, this seems highly unlikely. Note that this conclusion comes from our general understanding of how the real world works, NOT from the data itself.

A second important question is “Why did %P become HI over 69-90?” In connection with Friedman’s hypothesis, it is interesting to note that %M became high much later than these periods of high inflation. Money growth rates became consistently high in the period 79-91. Why did increase in money growth FOLLOW the increases in inflation? According to one theory, monetary policy should accommodate the needs of business. In periods of high inflation, the need for money is high, and so one should print more money. We need to look at the minutes of the Monetary Policy Committee to see what they were doing and why they were doing it. It seems clear that the periods of HIGH inflation from 1969 to 1979 were not preceded by High rates of growth of money (%M). It seems likely that this inflation came from a different source. Again, we need to more carefully at the real world, and try to find other causes of inflation, to explain the patterns we see in the data.

From this preliminary analysis, it is clear that Real Statistics leads us to ask different KINDS of questions. In Fisherian statistics, we start by ASSUMING that the Data are RANDOM draws from hypothetical population. In this case, the data is ONLY a means to discovering the parameters of this IMAGINARY population. All of the sophisticated mathematical machinery of inference and hypothesis testing deals with issues of how we can use the data to learn about the imaginary population from which the data is ASSUMED to be a random sample. In this scenario, if we know the parameters of the imaginary population, we don’t NEED the data!! The parameters provide us with COMPLETE information about the data set!! The Individual Data points DO NOT MATTER and are MEANINGLESS – they are all random draws and could come out differently next time. This is in dramatic contrast to real statistics: Each data point matters! We do not come to understand data by imagining a hypothetical underlying population. Instead, we understand Data by looking at the Reality which generates the data. Why did rates of money growth %M become high over 79-91? No amount of playing games with this data will lead us to the answers. Instead, we must examine the Australian Economy, and maybe the world economy as well. We must look at methods of money creation, to find sources for the extra money created. One source is the government. Look at bulletin of the Monetary Policy Committee. How were they making decisions about monetary policy. Were they accommodative or forward looking? What were the variables they considered? But additionally, we may study private money creation by financial institutions. This was a period of Financial De-Regulation. Removal of restrictions led increase in loans and creation of money. This may be reason why %M was high in this period.

On more technical note, we can also use this partitioning of data to create a simple measure of association, which does not rely on unverifiable assumptions about imaginary populations. We can simply DIVIDE the data into two halves – those with HI %M and Lo %M. Now look at the behavior of %P on each of these halves separately. If nearly all of the HIGH value of %P occur within the High %M data set, then the two series must be highly correlated. Doing this simple counting of the data gives us the following table of counts:

| %P vs %M | %M=LO | %M=HI |
|-------------|-------|-------|
| Lo %P | 20 | 9 |
| Hi %P | 9 | 21 |
| Total | 29 | 30 |

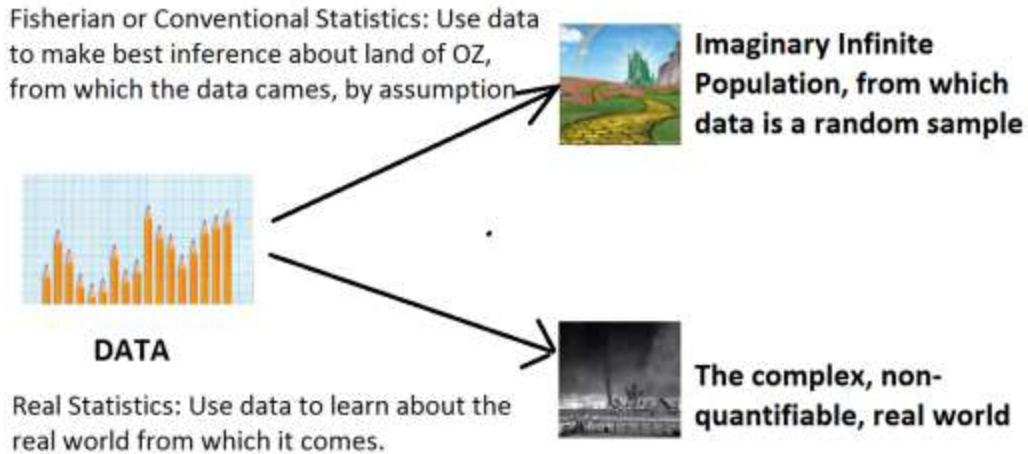
We see that in 29 years where the money growth was in the bottom half (%M is Lo), %P is also Low for 20 Years and High in 9 Years. Similarly, for the 30 years where %M is High, %P is high in 21 years and low in 9. This shows a strong association between the highs and lows of %M and %P. When one is high, the other one is also high in roughly 2/3 of the cases. One high and the other low occurs about 1/3 of the time. If there was no relationship between highs and lows of %M and %P, we would expect to see about 50% of the total or 15 cases out of 60, in each of the four boxes in the diagram above. The numbers show a strong but loose relationship. It seems clear that Friedman's hypothesis, the %M is the ONLY source of inflation, is not correct. While %M does exert a strong influence on %P, it seems likely that there are other causes of inflation as well – about 1/3 of the cases of Hi and Lo %P cannot be explained by Hi and Lo %M in the same period.

Conclusions

There is strong relationship between %M and %P. Direction of causation is not clear, and CANNOT be learnt from the data. INSTEAD, we can learn about causes by “expending shoe leather.” Expenditure shoe leather is a metaphor for exploring the real world and searching for causes. In this particular case, we must analyze bank statements and Monetary Policy statements, and look at real world factors for money creation and inflation.

We learn that the direct analysis of data, WITHOUT ANY assumptions about stochastic structure, gives us a lot of information about the real world. It is important to understand that NO MORE INFORMATION is available. Stochastic assumptions REDUCE importance of data, by mis-directing our attention from the data itself to a hypothetical imaginary infinite population, from which the data is assumed to be a random sample.

The lessons of this lecture are summarized in the picture below:



6: Five Quartile Summaries of Stochastic Relationships

A BLURB describing main points of this chapter.,

6A: Do the Wealthy Have Fewer Children?

asier goal. Various kinds of devices are used to get PROBABILITY samples – where you can calculate the probability of inclusion for everyone, even though it is not EQUAL. These are advanced topics, which we will study in later courses.

We note that the Computational Capabilities at our disposal to analyze the HIES data set were not available to Fisher. The simplest possible Fisherian Analysis starts by ASSUMING data is Bivariate Normal. This allows reduction of the TWO Variables – HH Size and HH Expenditure ($15,510 \times 2 = 31K$ Data Points) to FIVE numbers: sums of the series (2), sums of the squares of the series (2), and a sum of the product of the two series. It was part of the mathematical genius of Fisher to see that these 5 numbers would provide a complete summary of 31,000 data points IF the data was a random sample from a bivariate normal distribution. If the Fisherian assumption did not hold, there was simply no computational capability to analyze the entire data set! In fact, in Fisherian times, even computing these FIVE numbers from the HIES DATA Set would require several man-days of labor! The kinds of computations we will recommend were not even THINKABLE in those pre-computer days.

Even though it makes data analysis possible without a computer, there are serious problems with the Fisherian Assumption. On this particular data set, with HH S being an INTEGER from 1 to 60 – Normality CANNOT hold. Similarly HH Exp is highly SKEWED, clearly not a normal distribution. For such data sets, there was an attempt to find “data transformations” which would CREATE normality. For example, LOG (HH Exp) may be closer to normality, on this data set. However, ALL types of Fisherian analyses REPLACE the REAL

finite population with an IMAGINARY theoretical infinite population. The ONLY VIRTUE of this replacement is that it makes data analysis possible without computers. Unfortunately, the ASSUMPTION is BLATANTLY false – There is an ACTUAL population, with an ACTUAL distribution, which fails to match ANY of the theoretical distributions which are the subject of study in heavy statistical textbooks.

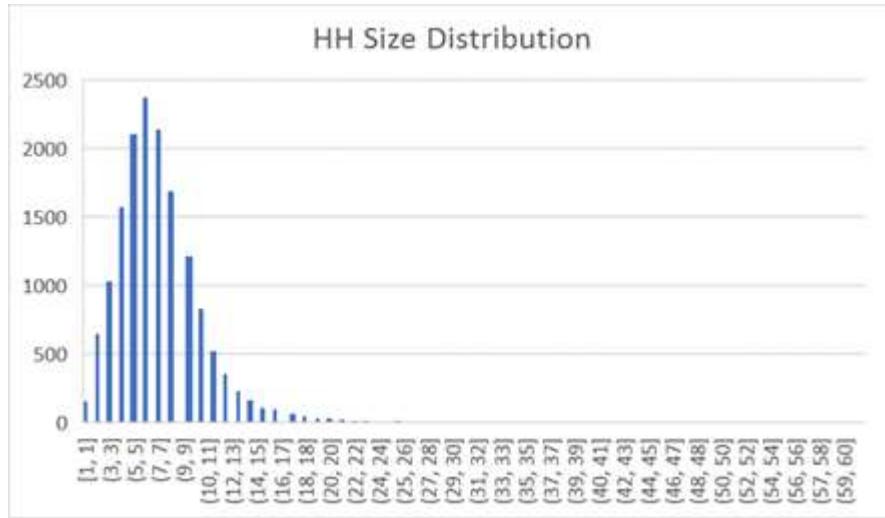
We are now light-years ahead of Fisher in terms of computational abilities. Whereas it would take multiple man-days to compute the sufficient statistics for this HIES data set of 15,509 x 2 data points, we can compute these and MUCH MORE with just a few keystrokes and clicks. Instead of ASSUMING a theoretical distribution for a hypothetical imaginary parent population, we can compute the EXACT distribution of the ACTUAL sample that we have. We will do so shortly. This distribution will have NO NAME – it does not belong to any theoretical family of parametric distributions. We cannot do fancy math with it. BUT WE CAN LOOK AT IT !!!

To begin with, let us illustrate a CRUDE Fisherian Analysis of the HIES Data Set. Assume the data to be bivariate normal – this is a theoretical distribution which has the simplest possible analysis. Calculate Means and Variances for both S,E, plus Covariance (S,E). These numbers are given in the table below, computed in seconds in EXCEL:

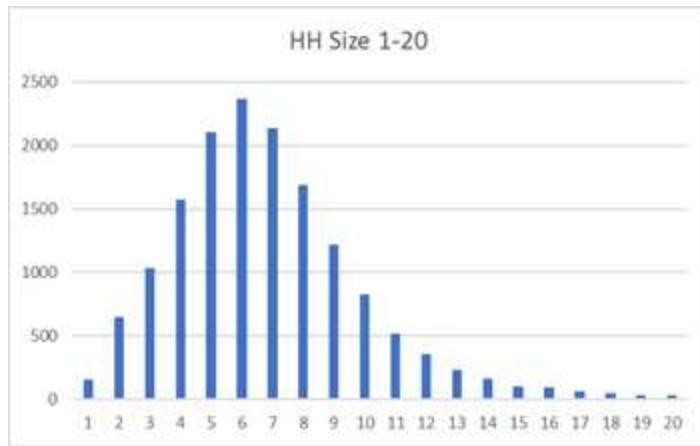
| | | | |
|---------------|------|---------|-----------------------|
| Avg HH Size= | 6.9 | 158,454 | =Avg HH Total Exp |
| Std Dev HH S= | 3.39 | 132682 | =Std Dev HH Total Exp |
| Correl= 0.252 | | | |

A traditional statistics course would make a major effort to teach students the meaning of the assumption of normality, as well as that of the average, the standard deviation, and the correlation. These concepts are essential for analysis of normal distributions, but do not work very well when the underlying distributions are not normal. Unlike conventional courses, we will not deal with normal distributions in this course, and with the associated concepts of mean, standard deviation, correlation. What is of interest from this classical Fisherian analysis is that there is a positive correlation between HH Size and HH TE (Total Expenditure). This means that as the HH TE increases, the HH Size also increases. But this contrary to the basic idea we are pursuing, according to which wealthy people have smaller families. The data seems to be telling us that wealth – as measured by HH Total Expenditure – goes together with HH Size. More of one means more of the other. This is puzzling – we will resolve this puzzle later.

We just presented a brief sketch of how a Fisherian analysis would go, without explaining the details, for the sake of illustration. Now we come to the heart of the lecture, the alternative methodology of Real Statistics. This is based on a direct analysis of data distribution. The latest versions of EXCEL allow us to produce a histogram – a picture of the data distribution – with a click. Here is the HH Size Histogram, as produced by EXCEL



HH Size Ranges from 1 to 60! We have a few very large households in Pakistan. However, there are very few households above 20; so few, that the bar is not even visible in the histogram. Nonetheless, showing the 20-60 portion of HH Size COMPRESSES the area available for the 1-20 range to only a 1/3 of the graph. This is harmful because this is where most of the data is. We can clarify the picture by Splitting the Graph into two different ranges 0-20 and 20-60. This is one of the main goals of Descriptive Statistics: to learn to LOOK at data. This simple lesson about how to make graphs which focus on the main parts of the data is NOT a part of traditional statistics, since traditional statistics is not concerned with the display of the data. However, this is an important lesson for descriptive statistics. Here is a graph of the first portion, with 0-20 HH Size, omitting the higher HH Sizes:



We can also convey this information in the form of a table which tells us, for each HH size, how many households in the sample of 15,509 have that size. Here is the table:

| Size | HH | # | HH | # |
|------|-----|------|-----|----|
| | HH | Size | HH | HH |
| 1 | 158 | 11 | 520 | |

| | | | |
|----|------|----|-----|
| 2 | 649 | 12 | 356 |
| 3 | 1032 | 13 | 235 |
| 4 | 1573 | 14 | 166 |
| 5 | 2109 | 15 | 103 |
| 6 | 2370 | 16 | 96 |
| 7 | 2139 | 17 | 64 |
| 8 | 1692 | 18 | 48 |
| 9 | 1217 | 19 | 34 |
| 10 | 829 | 20 | 30 |

This table gives us information about the HH Size for the HIES sample of 15,509 HH's. We see that the largest number of HH's – 2370 – have size 6. Sizes 5 and 7 are the next most popular with roughly 2100 HH's at each of the two sizes. Once we get to size 20 and above, only a very few HH's in the sample have such a large size. This is a direct look at the data, without any of the "SUMMARY STATISTICS" which are an essential part of conventional statistics.

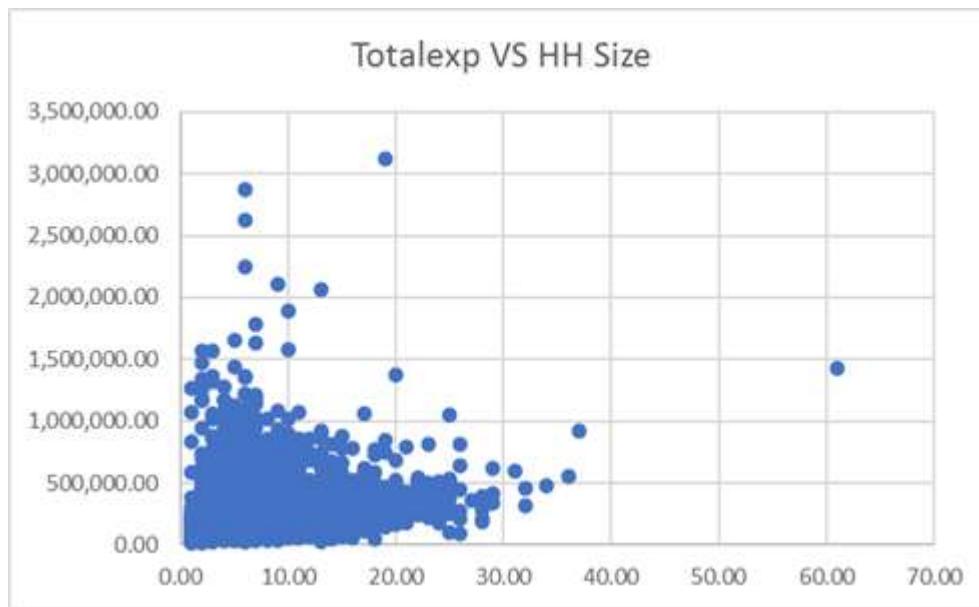
Just like the HH Size, we can also LOOK at the HH TE (Total Expenditure per Household). Note that this is a Luxury which was NOT available to Fisher. It would have required enormous number of man-hours to make the following graph of HH TE, which shows a Typical Income Distribution



This graph is HIGHLY skewed to the right, showing right away that it is not normal (Normal distribution is symmetric around center). Skewness means that there are a FEW people who are very, very, rich. The proportion is so small you cannot even see the bars on the graph.

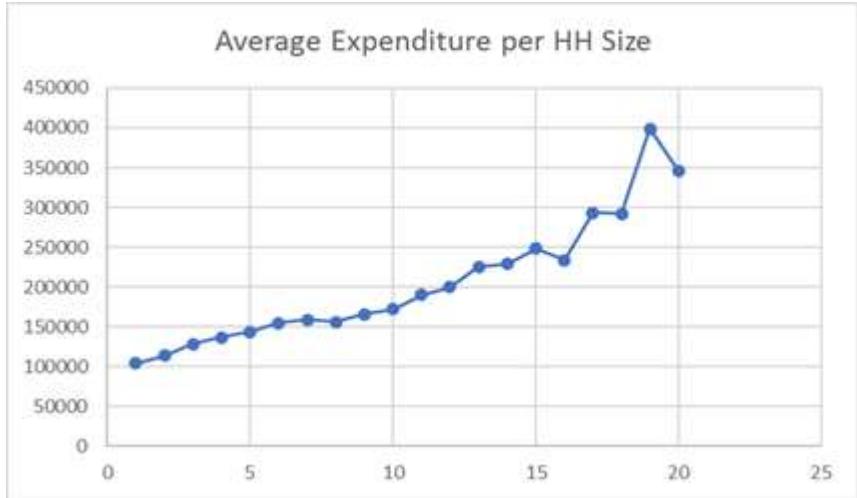
But since EXCEL draws the graph to go from the top to the bottom of the range, we can FEEL the presence of these super-rich people in the form of the long X-axis. As before, a better graph would result from SPLITTING into two parts; one part for the masses of people, and a separate one for the super-rich. Since this is not our main concern for the moment, we turn our attention to the relationship between the HH S and the HH TE.

This is main topic of interest – how does HH Size vary with wealth. We can simply ask EXCEL to produce an X-Y plot of these two variables. Each HH is plotted in the X-Y plane as a dot (X,Y) where X is the HH Size and Y is the HH TE. The graph looks as follows:



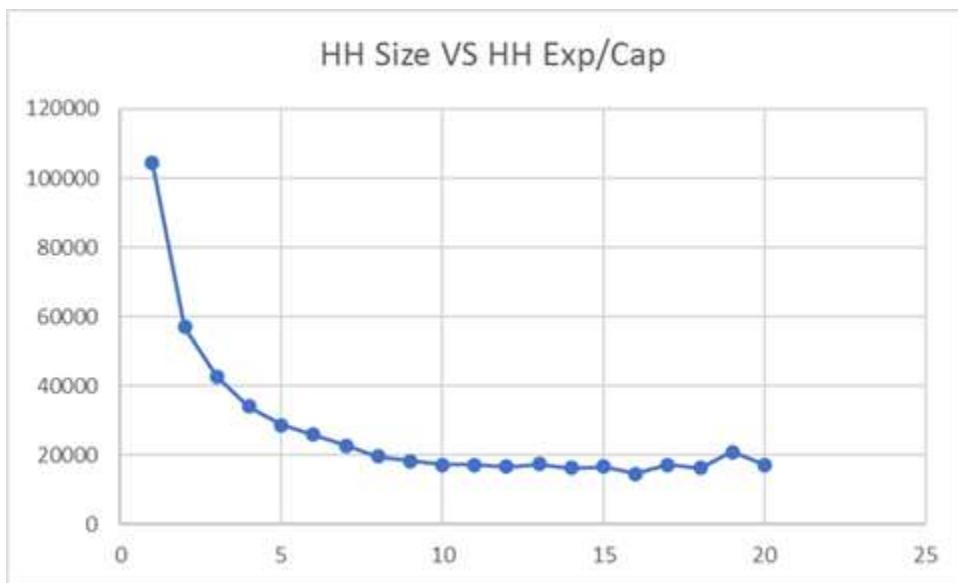
This is NOT a good way to look at the data. Thousands of points are plotted within a small space so we have no real idea of what the data looks like. The graph is too crowded. The reason we made it is because EXCEL can do it with a click. Also, to show the student the need of thinking about how to plot the data in GOOD ways. HOWEVER, even though it is a bad graph, it does NOT show what we were looking for. We thought that as wealth (represented by HH TE) increase, HH Size would go down – richer families have fewer children. But the graph shows a positive association. More HH TE goes together with bigger HH Size.

Another way of making the graph makes the picture even clearer, and also gives us some practice in making better graphs. Categorize the HH's by size. In each category, compute the AVERAGE total expenditure for ALL of the families in that category – that is, all HH with same HH Size. This can be done easily with EXCEL commands, and leads to the following result:



As the HH Size increases from 1 to 20, the average HH TE in that size HH ALSO increases, as the graph shows. This is in CONFLICT with the idea that rich families have FEWER children. Higher Average Expenditure corresponds to more children. However, intuition, experience, and general observations around the globe, tell us that wealthy families generally have fewer children. How can we resolve this MYSTERY?

The SOLUTION to this mystery comes when we realize that HH TE (Total Expenditure) is NOT the right PROXY for HH Wealth. Rather, we should use: HH TE per CAPITA – how much household spends on EACH person in the household. To see this, note that if HH Size 1 has TE=10,000 and HH Size 2 has TE=15,000, HH Size 2 is POORER, even though it has larger TE. This is because it spends 7,500 per person on two people, while HH 1 spends 10,000 on 1 person. We can get a better Proxy for wealth by dividing the Total Expenditure by the HH Size. Dividing HH TE by HH Size gives us HH TE per capita. We can take the new variable and recreate the same chart as in the previous picture – the AVERAGE Wealth (proxied by TE per capita) for each HH Size. This is plotted below:



FINALLY, this chart SHOWS what we want to see; as HH Size increases, HH TE/Capita declines. Larger Households have less money available for members on a per capita basis. Note the this decline is very steep in the early portion of the chart, while the curve becomes pretty flat after HH Size of 10 or so. The biggest effects are seen at the earliest parts of the curve, at low HH Size. Perhaps we could say that the wealthy households have 1,2, or 3 children but no more. The less wealthy can have any number of children. The graph just gives us the picture of the data, without telling us anything about the causal connections. We do not know if increased wealth leads to fewer children, or whether having more children depletes wealth, or whether there is some third factor we have ignored, which creates this apparent relationship between wealth and family size.

CONCLUSIONS

So what do we learn by this preliminary REAL data analysis of the 2006 HIES data set? We learn that direct examination of the data – made possible by EXCEL — gives us a lot of information, WITHOUT any statistical assumptions. In contrast, Fisherian Statistics TYPICALLY involves IMPOSING assumption of Multivariate Normality on Data. When this assumptions is made, then Means, Standard Deviations, Correlations, COMPLETELY summarize data. This was one of the brilliant mathematical contributions of Fisher – he proved that these five numbers are SUFFICIENT STATISTICS – once you have them, you can throw away the data without loss of information.

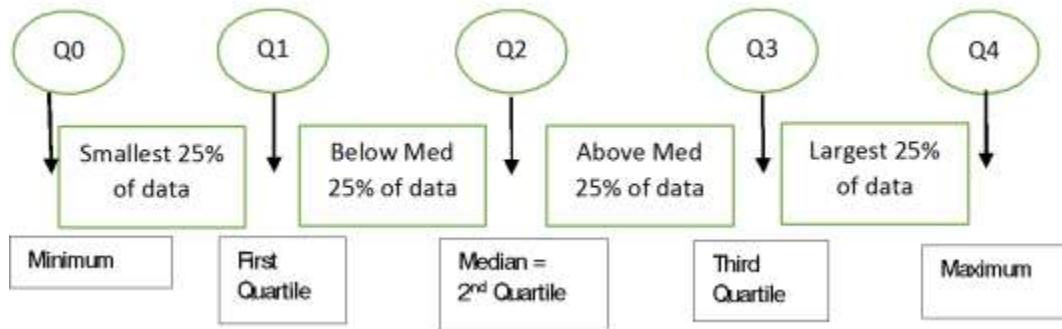
Of course, this sufficiency was based on the ASSUMPTION that the theoretical distribution assumed is VALID. Fisher was ASKED about how you choose a statistical model; the theoretical distribution for the data. He DODGED the problem by referring to the practitioners, saying that they would know, from their practical field experience, the right theoretical model. For a century, the entire field of statistics has been built on the idea that the first thing to do is to ASSUME a theoretical statistical model for the data. This is because such an assumption allows the reduction of the data to a few sufficient statistics, and enables analysis. With our current computational capabilities, we NO LONGER need to make arbitrary assumptions to reduce the data. We are now capable of dealing with the FULL DATA set of 15,509 points, without trying to reduce it. However, statistician have invested a century of effort in developing tools for data reduction, and they are not likely to give up using them anytime soon. Neils Bohr, inventor of Quantum Mechanics noted, in frustration at the reluctance of physicists in accepting his theories, the Physics progresses one funeral at a time.

The REAL statistics we propose is to do is nowhere as complex as Quantum Mechanics. Quite the opposite, we propose to eliminate complex mathematical assumptions and analysis. Real Statistics starts just by LOOKING at the data in intelligent ways, using graphs and tables, whatever type of method is suitable to reveal aspects of reality we are exploring. This can only be learnt in an apprentice like fashion, by studying case after case of intelligent visualization of the data. This is our hope for this course.

6B: Quartiles as Natural Data Summaries

A Fisherian approach to statistics begins by ASSUMING that the data is a random sample from a theoretical population characterized by a small number of parameters. Such an assumption has no basis in reality, but is made to make statistical computations possible in a pre-computer era. Because of inertia, this century-old methodology continues to dominate the field, even though advances in computational capabilities have made it obsolete. Instead of an assumed distribution, REAL statistics take the ACTUAL data distribution as the central tool for data analysis. This actual distribution never belongs to any of the neat theoretical families of distributions that make for elegant mathematical analysis. It cannot be written down on paper as a formula, but it can easily be computed by the computer. A visual depiction of the actual data distribution is the Histogram, which we have studied earlier. A more rigorous mathematical approach would be based on the Empirical Cumulative Distribution Function (ECDF) which we will study and explore later. The ECDF is directly based on the data, depends on all of the data, and does not allow for data reduction, unlike the Fisherian approach. It is just that we can now handle computations on 15509 data points in the HIES data with a click, so we do not NEED to reduce the data before we start the analysis.

It is nonetheless useful to have a few SUMMARY statistics which describe the data distribution. Our main concern in this lecture is to develop the concept of the QUARTILES, as a natural and intuitive description of the data set. Parenthetically, we note that in the Fisherian approach, these summary statistics are the mean and the standard deviation, which work wonderfully for the hypothetical normal distributions, but are extremely poor for other distributions. Here is a visual description of how we define the Quartiles.

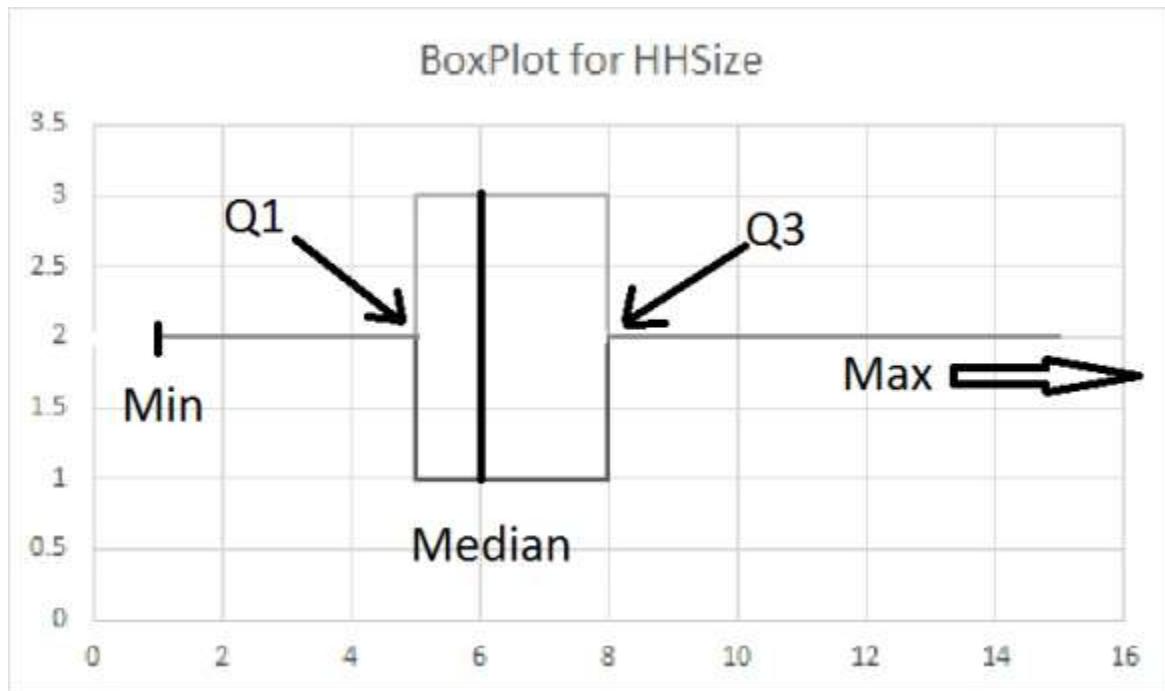


First we sort the data, so that it is arranged in increasing order. Then we divide the data into FOUR equal parts. With 15509 points of data, this comes out to about 3877 data points in each of the four portions. The Summary Statistics are the separating points for these four portions of the data: Q0, Q1, Q2, Q3, Q4. Note that HH Size is always an integer and varies from a minimum of 1 to a maximum of 61 on this data set of 15509 points. The summary statistics are computed as follows:

- $Q_0 = \text{Minimum HHS}$
- $Q_1 = \text{HHS}(3877) = 1\text{st Quartile}, 3877.25 = 15509/4$
- $Q_2 = \text{Median} - \text{HHS}(7754), 7754.5 = 15509/2$

- $Q3=8=HHS(11631)=3rd\ Quartile,\ 11631.73=3*15509/4$
- $Q5=61=Maximum$

We, humans, are much better at absorbing information in pictures and graphs, as opposed to numbers. That is why a Box-and-Whiskers plot (short form: Boxplot) provides a Graphic View of Summary Stats:



The LEFT whisker of the box-and-whisker plot goes from Q0 to Q1, the minimum to the first quartile. For HH Size the minimum value is 1, while the first quartile occurs at HH Size = 5. The third quartile is at HH Size = 8. The BOX is made between Q1 and Q3 and represents HALF of the data. 25% of the data in the lower whisker, while another 25% is in the RIGHT whisker which goes from Q3 to the maximum HH Size of 61. A line is drawn in the middle of the box to show where Q2 or the median belongs. Thus, all 5 quartiles Q0, Q1, Q2, Q3, and Q4 are pictured in the boxplot.

So what do we learn from this boxplot? Of the greatest importance is the Central Value or the median, which is HH Size = 6. What exactly does this mean? It means that HALF of the households have HH Size ≤ 6 , while HALF of the households have household sizes ≥ 6 . Thus the HH Size of 6 divides the population into two equal halves, where one half is smaller and the other half is larger. Some technical issues arise because HH Size is integer-valued and jumps from 5 to 6 to 7 without taking any values in the middle. Thus, when we look at HH Sizes of 1,2,3,4,5, less than 50% HHs these sizes. When we add the size 6, then more than 50% of the households have size 1-6. This is a technical issue that is not of importance for us in the present context.

After the CENTRAL VALUE or the median, the next most important thing is the SPREAD of the data, which is measured by the Interquartile Range. This is defined as the distance between Q3 and Q1. In this data set, the BOX goes from HH Size 5 to HH Size 8. This means that 50% of the households have sizes in the range 5,6,7,8. 25% or less have HH Size below {1,2,3,4}, while 25% have HH Size above {9,10,...,61}. This tells us that the distribution is Asymmetric; it is Right SKEWed. The left whisker is very short, so the data distribution has a Short Left Tail. On the other hand, it has an Extremely Long Right Tail. To understand the quartiles better, we show how we can compute them from the following table. For each HH Size, the table COUNTS the number of HouseHolds with SMALLER HH Size. Thus, the first entry shows that there are 3412 HH's which have size {1,2,3,4} (less than 5). We note that $3412/15,509 = 22\%$, so this is less than a quarter of the population. However, when we go to the next entry, that is 5521 HH's of size {1,2,3,4,5} and this is 35% of the population. So HH Size = 5 goes from 22% to 35.6% which COVERS 25% or the first quartile. Similarly, HH Size = 6 takes us from 35.6% to 50.6%, which COVERS 50% or the second quartile. Similarly, HH Size = 8 takes us from 64.7% to 75.6%, which COVERS Q3 = 75%.

| HH Size | < | % |
|------------|-------|--------|
| 5 | 3412 | 22.00% |
| 6 | 5521 | 35.60% |
| 7 | 7891 | 50.90% |
| 8 | 10030 | 64.70% |
| 9 | 11722 | 75.60% |

These problems, where the percentiles jump from 22% to 35.6% without coming close to 25% arise because HH Size is an integer and can only take certain fixed values. We next look at the Summary Stats for HH TE/cap (Total Expenditure per capita), which is a continuous variable. As we will see, these problems do not arise for continuous variables. A table similar to the one above lists the five quartiles of TE/cap in the first column. The second column list the NUMBER of HH's which have smaller TE/cap, while the 3rd column displays this number as a percentage of 15509.

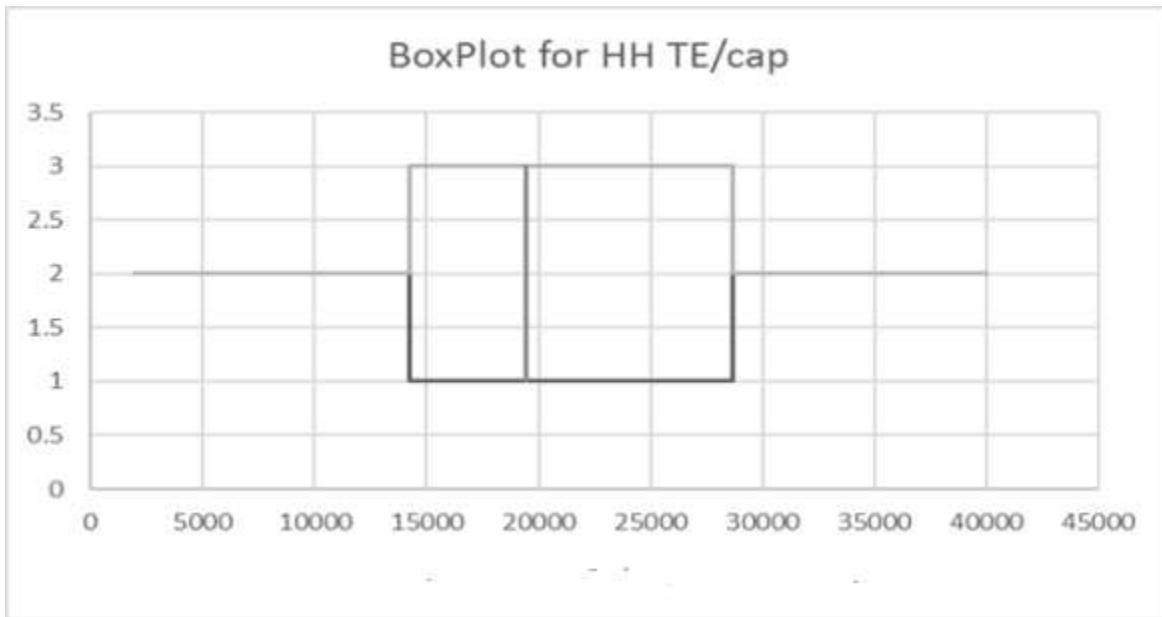
| TE/cap | #HH below | % HH below |
|------------|-----------|------------|
| MIN = 1966 | 0 | 0% |
| Q1 = 14275 | 3877 | 24.998% |
| Q2 = 19454 | 7754 | 49.997% |
| Q3 = 28648 | 11631 | 74.995% |

MAX = 1268708

15508

100%

A visual depiction of these quartiles can be seen in a boxplot:



The central value is the Median TE/cap = 19,454. This is central because 7754 HHs are below (having less TE/cap) and also 7754 are above, having more TE/cap. In traditional statistics, one might use the Average value of PKR 27,119 for the Center of this distribution. This is great if the distribution is normal, but it becomes Very Distorted due to the presence of huge outliers, which are not part of any normal distribution. In general, the widely used summary statistics of the Mean is best for Normal, but VERY BAD for general distributions. In contrast, the Median works well for ALL distributions and has a natural and intuitive interpretation.

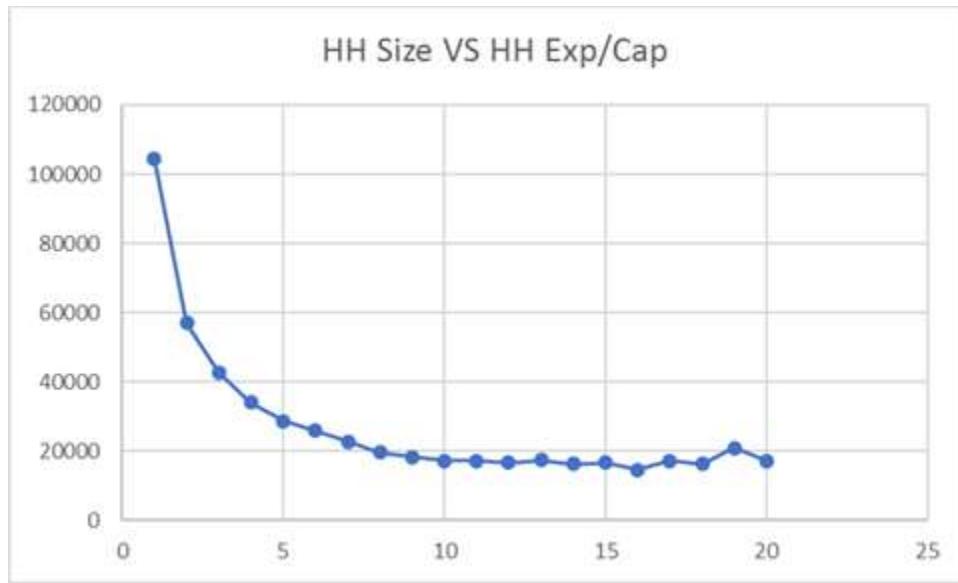
The next thing we learn from the data is the Dispersion: How Spread Out is the Data? The boxplot used the middle 50% of the data to measure this. The Interquartile Range. [14275, 28648] – Half of the households have TE/cap within this range. 25% have LESS and 25% have more. $IQR = 28468 - 14275 = 14193$. This is a natural measure of dispersion for general distributions. It is the REPLACEMENT for Standard Deviation which works well ONLY for normal distributions.

The boxplot also tells us about the Skewness & the Tails. Both HHSIZE and HH TE/cap are right-skewed. TE/cap is much more skewed. Both have large right tails — HHSIZE goes up to 61 – HH TE/Cap goes to 1,268,708. TE/cap has a much more extreme extension in the right tail. In contrast, the Normal distribution is symmetric and has thin tails.

We would like to study the relationship between HH Size and HH TE/cap (which is a proxy for HH Wealth). In conventional statistics, the methodology of doing this is based on “regressions”. As usual, these regressions are based on large numbers of unverifiable and false assumptions. Famous statistician David Freedman said that ‘we have been running regressions for a century. This has not led to any useful results. Let us abandon the technique’. In real

statistics, we propose to use the Median Line of X given Y as a REPLACEMENT for regression lines. We will illustrate this by drawing the two Median-lines, one of HH Size against TE/cap and the other for TE/cap against the HH Size. Intuitively, the idea is to create small boxes (bins) for one the variables, say Y. That amounts to making the range of variation small for that variable. Within a bin, the variable Y does not vary much. Now compute the MEDIAN value of X in this bin. That will tell us the central value of X for Y's within a particular box or bin. Now, as we change the Y-values, moving up across the Y-bins, we will find 203 949 5500 how the median of X changes as Y changes across bins. This will give us an idea about how the variable X responds to changes in the variable Y. We now illustrate the concept of Median-Lines for our HIES data set.

Conceptually, it is easier to see how the Median TE/cap varies with household size. We simply subdivide the data into groups according to HH Size. For each HH Size, we look at ALL the HHs with that size. Here is the Median Graph of TE/cap according to HH Size:



The first point on the graph shows that when HH Size = 1, the MEDIAN TE/cap is above 100,000. Note carefully what this means. It does not mean that ALL Households of size 1 are rich. Rather, there are (only) 159 HH's of size 1 in the entire sample of 15,509. Among these 159 HH's the median income is above 100,000 – that is more than half, or 80+ HH's, have income in excess of 100,000. 80 of the HHs in this group (having size = 1) have TE/cap LESS than 100,000. Similarly, for each category of HH Size, the dot shows the median TE/cap of all HHs having that size.

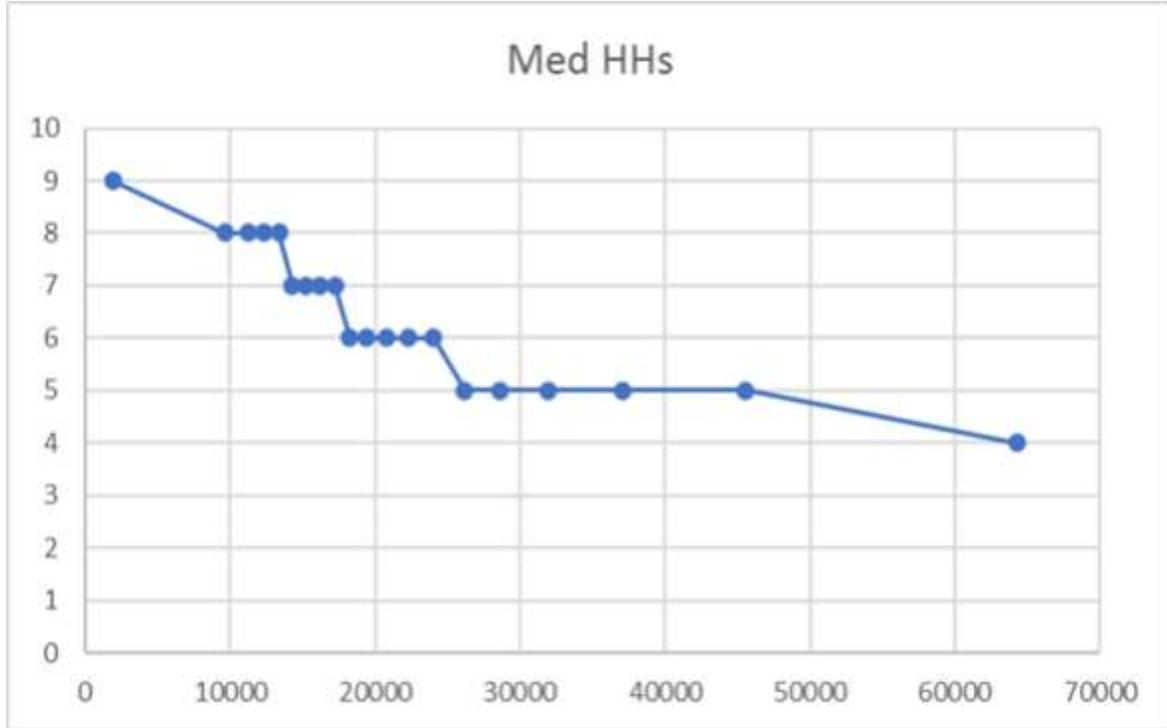
It is clear from the graph that, as HH Size Increases, MEDIAN TE/cap decreases. The most rapid changes occur early, for small HH Size. From HH Size of 1 to 5, there is rapid reduction of Median TE/cap as HH Size increases. From HH Size 5 to 10, there are small reductions in median income. After HH Size = 10, the Median line seems pretty flat.

Next, we consider the other Median-LIne of HH Size for TE/cap groups. In order to create this, the first step is to subdivide TE/cap into small buckets. There are many possibilities,

but in the present case, a natural method is as follows. We note that $775 \times 20 = 15500$, so if we create 20 buckets, with each bucket having 775 families, we will cover 15500 families. To cover the remaining 9 families, we can just add one family to every other bucket. We will describe the full technical details how to do these operations in EXCEL in the next portion of this lecture. For the moment, we just note the income groups which are created by this procedure are as follows:

| Gr oup | Gr oup No. | TE/ cap Lo | TE/ cap Hi | Gr oup Size | Gr oup | Gr oup No. | TE/ cap Lo | TE/ cap Hi | TE/c ap | Gr oup Size |
|-----------|------------------|------------------|------------------|-------------------|-----------|------------------|------------------|------------------|------------|-------------------|
| 1 | 6 | 196 | 9 | 960 | 775 | 11 | 55 | 194 | 9 | 2081 |
| 2 | 4 | 961 | 20 | 112 | 776 | 12 | 21 | 208 | 8 | 2230 |
| 3 | 21 | 112 | 24 | 123 | 775 | 13 | 09 | 223 | 8 | 2401 |
| 4 | 24 | 123 | 59 | 133 | 776 | 14 | 24 | 240 | 1 | 2615 |
| 5 | 59 | 133 | 75 | 142 | 775 | 15 | 54 | 261 | 8 | 2864 |
| 6 | 75 | 142 | 35 | 152 | 776 | 16 | 51 | 286 | 9 | 3190 |
| 7 | 36 | 152 | 99 | 161 | 775 | 17 | 16 | 319 | 8 | 3709 |
| 8 | 00 | 162 | 35 | 172 | 776 | 18 | 00 | 371 | 5 | 4556 |
| 9 | 36 | 172 | 72 | 182 | 775 | 19 | 76 | 455 | 7 | 6429 |
| 10 | 73 | 182 | 54 | 194 | 776 | 20 | 17 | 643 | 708 | 1268 |

Each of these 20 buckets has 775 or 776 families. Now we look at each of these buckets separately, and compute the MEDIAN HH Size for each group of 775/776 families. These can be plotted as follows



This is a graph of the Median HH Size for each of the 20 income groups as described above. This same information can be given in tabular form as follows:

| TE/cap HHs | Med HHs | TE/cap HHs | Med |
|---------------|------------|---------------|-----|
| 1966 | 9 | 19455 | 6 |
| 9614 | 8 | 20821 | 6 |
| 11221 | 8 | 22309 | 6 |
| 12324 | 8 | 24024 | 6 |
| 13359 | 8 | 26154 | 5 |
| 14275 | 7 | 28651 | 5 |
| 15236 | 7 | 31916 | 5 |
| 16200 | 7 | 37100 | 5 |
| 17236 | 7 | 45576 | 5 |
| 18273 | 6 | 64317 | 4 |

Both the graph and the table provide us with the same information. As we go up the TE/cap groups, the median HH Size decreases. This supports the idea that wealthier families have fewer children. But it also supports the reverse causality. That is, having more members in a Household reduces the amount of money available per member. That is, large HH Size leads to poverty. Understanding causality is of essential importance, but this cannot be learned from the data – the data does NOT provide the information required to learn about the causal directions.

Conclusions

We have done this data analysis without any assumptions about randomness. Even though we are using the word “distribution” to describe the data, this is just an observed pattern that the data follow. We are NOT making any assumption that the data is a random draw from any distribution at all. Fisherian old-school statisticians will find this terminology very confusing, because we are using similar words with different meanings. For example, the Median-Lines a description of the “conditional distributions” of HH TE/cap given HH Size and also of HH Size given HH TE/cap. More discussion of this subtle issue will be given later in the course.

Both Median Lines show that wealthier families have fewer children – conversely, small HHs correspond to higher TE/cap. Note that variables have been CAREFULLY chosen – This result holds for TE/cap but not for TE. Real Statistics requires relating data series to real concepts, not just treating them as numbers. Our Median-Lines show ASSOCIATION between the two variables. CAUSALITY cannot be learned directly from the data. The last point of great importance is that the relationship between HH Size and HH TE/cap is not deterministic. At any given HH Size, we have a large range of Households with very different TE/caps. Similarly, in every income (TE/cap) group, there is a large range of HH Sizes. How to understand these “flexible” relationships, also called “stochastic” relationships, will be the subject of the next portion of this lecture.

6C Stochastic Relationships

In this lecture we will discuss the concept of a *stochastic relationship*. For this purpose, a New Conceptual Framework is Required. We are used to thinking in terms of Deterministic Causal Relationships – I do X, it causes Y. Stochastic relationships are different. To explain by an example, suppose that when patients get disease D, 50% of patients recover on their own, without any treatment. Now a drug is invented, such that with the Drug, 90% of patients recover. The drug IMPROVES the recovery rate, but does not guarantee recovery. Given 1000 sick patients, 500 of them will recover without any treatment. If we administer the drug to all 1000, then 900 of them will recover, but 100 will still die. Two types of common-sense objections to stochastic relationships are OFTEN made by the public:

1. Drug does not matter for recovery, because 500 patients (out of 1000) recovered WITHOUT Drug.
2. Drug does not matter because 100 patients (out of 1000) DID NOT recover even though they took the drug.

Note that whenever we look at a case-by-case basis, we will not be able to SEE a stochastic relationship. Yes, there are patients who took the drug and did not recover. MANY patients recovered without the drug. The efficiency of the drug can ONLY be established by looking at the whole group together, BOTH the treatment group who took the drug, and also the control group which did not take the drug. A stochastic relationship is NOT between X & Y, BUT between X and Distribution of Y. You can only learn about it by looking at how the DISTRIBUTION of Y changes as X changes, not be looking at individual cases of what happens to Y when X changes. We will look at some specific cases to develop a better understanding of this concept.

In the previous lecture (dsia06b), we discussed the Median Line. This shows how the MEDIAN of the distribution of Y changes as X changes. This is one important aspect of how the distribution of Y changes, but the distribution of Y is much more than the median of this distribution. The QUESTION of “WHAT is the distribution of Y?” becomes much more urgent and important to answer, in this context. In conventional Fisherian statistics, this question has a clear and easy answer. We ASSUME the data is a random sample from a hypothetical theoretical distribution characterized by a few parameters. This theoretical distribution IS the distribution of the data. BUT this is not satisfactory from a REAL statistics perspective – this parametric distribution is ONLY an assumption and has no direct contact with the data or reality. In REAL Statistics, we take the distribution of the data to be the actual observed data distribution. There is a technical way to define this distribution which we provide below for the sake of completeness, even though we will not use this definition in this basic course. It will become important when we go to more advanced concepts.

DEFINITION OF A RANDOM SAMPLE: X is a random sample from a population P, if it is chosen in such a way that ALL members of P have an equal chance.

DEFINITION OF THE DATA DISTRIBUTION: Let X be a random sample from the DATA. That is, choose ONE data point at random, in such a way that ALL of the data points have an equal chance of being chosen. The distribution of random variable X is the DATA distribution.

This course is meant as a first course in statistics, for students who have no previous familiarity with the concepts of random samples, random variable, and distributions. The above definitions are given just for completeness, and will not be used in the material that follows. Coming from the theoretical background to the practical, the graphical form of the data distribution is the data histogram, a concept that we have studies in detail. The mathematical form is the Empirical Cumulative Distribution Functions (ECDF), which we will define later, in conjunction with deeper mathematical concepts. For the purpose of our basic course, we are more interested in the issue of: “What does it MEAN?” This question become especially critical in absence of Fisher-type assumptions. In conventional courses, we START by ASSUMING the data is NORMAL, so the data distribution is normal by assumption. In REAL Statistics, we start out with NO assumptions about the data. The meaning of the data distribution can be different for different types of data. The best way to study these meanings is by the apprenticeship

method, where we study the meaning on a case-by-case basis. For each REAL data set, the data distribution has a particular fixed meaning with reference to the real-world origins of the data.

Because the entire distribution is difficult to analyze, it is useful to concentrate on the SUMMARY Statistics, in particular, the five quartiles, of the distribution. Then we can address the question of “How does DISTRIBUTION of Y change as X changes?” by analyzing how the summary statistics of the distribution change. ONE of these summary statistics is Q2 or the median. The MEDIAN Line: How the MEDIAN of the distribution of Y changes with X. We saw that as TE/cap increase Median HH Size decreases. The graphic from this previous lecture discussed this earlier.

In addition to the median, a second important TOOL of real statistics is the MIDRANGE of the data distribution. This is a NEW use of the word. The Old meaning: $(Q_0+Q_5)/2 = (Min+Max)/2$ – also called Mid-Extreme. This concept is rarely used for a number of reasons. It is convenient to propose a New Meaning for the word MIDRANGE: The MIDDLE HALF of the data set. This is the collection of points between Q1 and Q3, which consists of HALF of the data. Below Q1 we have 25% of the data while above Q3 we have the top 25% of the data. We will now ILLUSTRATE the concept by its use on a real data set. This is important aspect of METHODOLOGY of REAL Statistics. We must link theory to practice. When we introduce a theoretical concept, we explain its meaning and use within the context of a real world example.

To illustrate a stochastic relationship, we continue our analysis of the HIES data set example. We will show how the Quartiles of the distribution of Y change as X changes where X and Y are HH Size and TE/cap the two variables in the HIES data set. We will also provide some Technical Details about how we make these Quartile Lines in EXCEL.

We start with the original Data in the first two columns of an EXCEL Spreadsheet. HH Size in A1:A15510, TE/Cap B1:B15510. Note that the columns are labelled, so the first row is the NAME of the data series, which means that there are 15,509 points of data in each column (plus one entry for the name). Here is a view of the first few columns of the EXCEL Spreadsheet:

The screenshot shows an Excel spreadsheet titled "DSIA06B Quartiles.xlsx". The "Home" tab is selected. The formula bar at the top shows "C6". The table below has four columns: A, B, C, and D. Column A contains values 1 through 9. Column B contains values 13, 18, 2576, 3024, 16, 26, 9, 13, and 25. Column C contains values 6.91186, "", "", "Min", "Q1", "Q2", "Q3", "Max", and "3998". Column D is empty.

| | A | B | C | D |
|---|--------|---------|---------|-----|
| 1 | HHSize | Exp/Cap | 6.91186 | |
| 2 | | 13 | 1966 | |
| 3 | | 18 | 2576 | |
| 4 | | 14 | 3024 | Min |
| 5 | | 16 | 3394 | Q1 |
| 6 | | 26 | 3603 | Q2 |
| 7 | | 9 | 3840 | Q3 |
| 8 | | 13 | 3890 | Max |
| 9 | | 25 | 3998 | |

This data set has been modified from the original. The HH Size is same as in original data set. However, we have replaced Total Expenditure of Household (HH) by Exp/Cap which comes from dividing Total Expenditure by the HH Size, giving the expenditure per member of the HH. Also, the data set above has been sorted by Exp/Cap. The above diagram also displays the five summary statistics – Q0, Q1, Q2, Q3, Q4 – for the two variables. These can be computed via simple EXCEL commands as listed below:

| | HHSIZE | TE/CAP |
|-----|----------------------------|----------------------------|
| Min | =MIN(A\$2:A\$15510) | =MIN(B\$2:B\$15510) |
| Q1 | =QUARTILE(A\$2:A\$15510,1) | =QUARTILE(B\$2:B\$15510,1) |
| Q2 | =QUARTILE(A\$2:A\$15510,2) | =QUARTILE(B\$2:B\$15510,2) |
| Q3 | =QUARTILE(A\$2:A\$15510,3) | =QUARTILE(B\$2:B\$15510,3) |
| Max | =MAX(A\$2:A\$15510) | =MAX(B\$2:B\$15510) |

These commands use MIN and MAX for Q0 and Q4, and the Quartile Command for Q1,Q2,Q3, and apply them to the entire data set in one of two columns to get the five quartiles as summary statistic for the entire data set. NEXT, we want to look at the summary statistics in subgroups of the data where one of the two variables is kept confined to a small interval (to keep it nearly constant). That way, we can study the change in distribution of the other variable, while keeping one of them fixed within an interval of values. In the previous lecture, we studies how Q2, or the MEDIAN, changes as the other variable changes. This was called the Median Line. In this lecture, we want to look at the range between Q1 and Q3, which contains half of the data. We will call this the MIDRANGE of the data (or, alternatively, the MIDDLE-HALF of the data). We now show how we compute and graph this midrange in EXCEL.

First, we will compute the Mid-Range of Exp/Cap for each HH Size separately. The FIRST STEP towards this goal is to SORT the data by HH Size. To do this, click on the cell A1. Then we must First HIGHLIGHT the data set (A1:B15510). – Then use SHIFT ® to highlight A1 and B1. Next, use CTRL-SHIFT- - to highlight the entire data set. Then click on *Sort and Filter*. If you click on the first menu item (- Sort A to Z), EXCEL will sort the data in increasing order according to the first column, which is HH Size. This will automatically put the HH Size in increasing order, starting with HH Size = 1 and going up to 2,3, etc. It is of importance to note that the SECOND column will also get sorted within the same HH Size. That is, if we look at the 158 HH's with size one, they will be arranged in order of increasing Exp/Cap. Then we start all over for HH Size = 2. There are 649 HH's with size 2, and they are again arranged in order of increasing Exp/Cap. In a stochastic relationship, we want to look at the DISTRIBUTION of the Exp/Cap of the 158 HH's of Size 1, and then compare with the distribution of the Exp/Cap of the 649 HH's of size 2. Similarly, as we go up in HH Size, we get different groups of HH's in the sample. How the DISTRIBUTION of Exp/Cap changes in these groups gives us the STOCHASTIC relationship between HH Size and Exp/Cap.

For the technical details of how we do with this EXCEL, we proceed as follows. The first step is to determine the size of group – how many households of each size are there? Remember that the first column (A2 to A15510) contains the household sizes. The EXCEL command =COUNTIF(A2:A15510,”=1”) counts all the entries in the cells A2 to A15510 which are equal to 1. Replacing “=1” by “=2” counts all the entries equal to 2, and so on. Using these commands, we can get the count of the HH of each size, which is displayed below:

| HHSize | Count | HHSize | Count |
|--------|-------|--------|-------|
| 1 | 158 | 11 | 520 |
| 2 | 649 | 12 | 356 |
| 3 | 1032 | 13 | 235 |
| 4 | 1573 | 14 | 166 |
| 5 | 2109 | 15 | 103 |
| 6 | 2370 | 16 | 96 |
| 7 | 2139 | 17 | 64 |
| 8 | 1692 | 18 | 48 |
| 9 | 1217 | 19 | 34 |
| 10 | 829 | 20 | 30 |

As we can see, the number of HHs falling into each category increases until HH Size 6, and then it decreases. At HH Size 20 there are only 30 families among the entire 15509 which have HH Size 20 (presumably due to extended families). Even though HH Size goes up to the maximum value of 61, it is useful to cut off the high value so that our graph gives more space to the smaller values where there is the most amount of data. The total number of HHs of sizes 1 to 20 is 15420 (from adding up all of the entries above), which about 90 less than the total data set of 15509. So ignoring the HH Sizes above 20 only misses a small number of HHs.

The next step is to find the quartiles of the Exp/cap in each of these 20 groups of HHs with fixed HH Size. For example, the first group with HH Size = 1 has 158 members. This means that the cells B2:B159 contain the Exp/Cap for this group. The next group with HH Size = 2 has 649 members. Thus, the cells from B160 to B808 (=159+649) contain the Exp/Cap for the second group. In a similar way, we can calculate the array of Exp/Cap cells for each HH Size. This is given below:

| H H Size | H #H | s Start | Ends | H H Size | H #H | s Start | Ends |
|-------------|---------|------------|------|-------------|---------|------------|------|
| 1 | 159 | B2 | B159 | 1 | 142 | B137 | B142 |
| 2 | 808 | B160 | B808 | 2 | 146 | B142 | B146 |

| | | | | | | | | | | | | | | |
|---|----|-----|----|------|----|------|---|---|----|-----|----|------|----|------|
| 3 | 0 | 184 | | B809 | 0 | B184 | 3 | 1 | 80 | 148 | 46 | B146 | 80 | B148 |
| 4 | 3 | 341 | 1 | B184 | 3 | B341 | 4 | 1 | 46 | 150 | 81 | B148 | 46 | B150 |
| 5 | 2 | 552 | 4 | B341 | 2 | B552 | 5 | 1 | 49 | 151 | 47 | B150 | 49 | B151 |
| 6 | 2 | 789 | 3 | B552 | 2 | B789 | 6 | 1 | 45 | 152 | 50 | B151 | 45 | B152 |
| 7 | 31 | 100 | 3 | B789 | 31 | B100 | 7 | 1 | 09 | 153 | 46 | B152 | 09 | B153 |
| 8 | 23 | 117 | 32 | B100 | 23 | B117 | 8 | 1 | 57 | 153 | 10 | B153 | 57 | B153 |
| 9 | 40 | 129 | 24 | B117 | 40 | B129 | 9 | 1 | 91 | 153 | 58 | B153 | 91 | B153 |
| 0 | 1 | 137 | 41 | B129 | 69 | B137 | 0 | 2 | 21 | 154 | 92 | B153 | 21 | B154 |

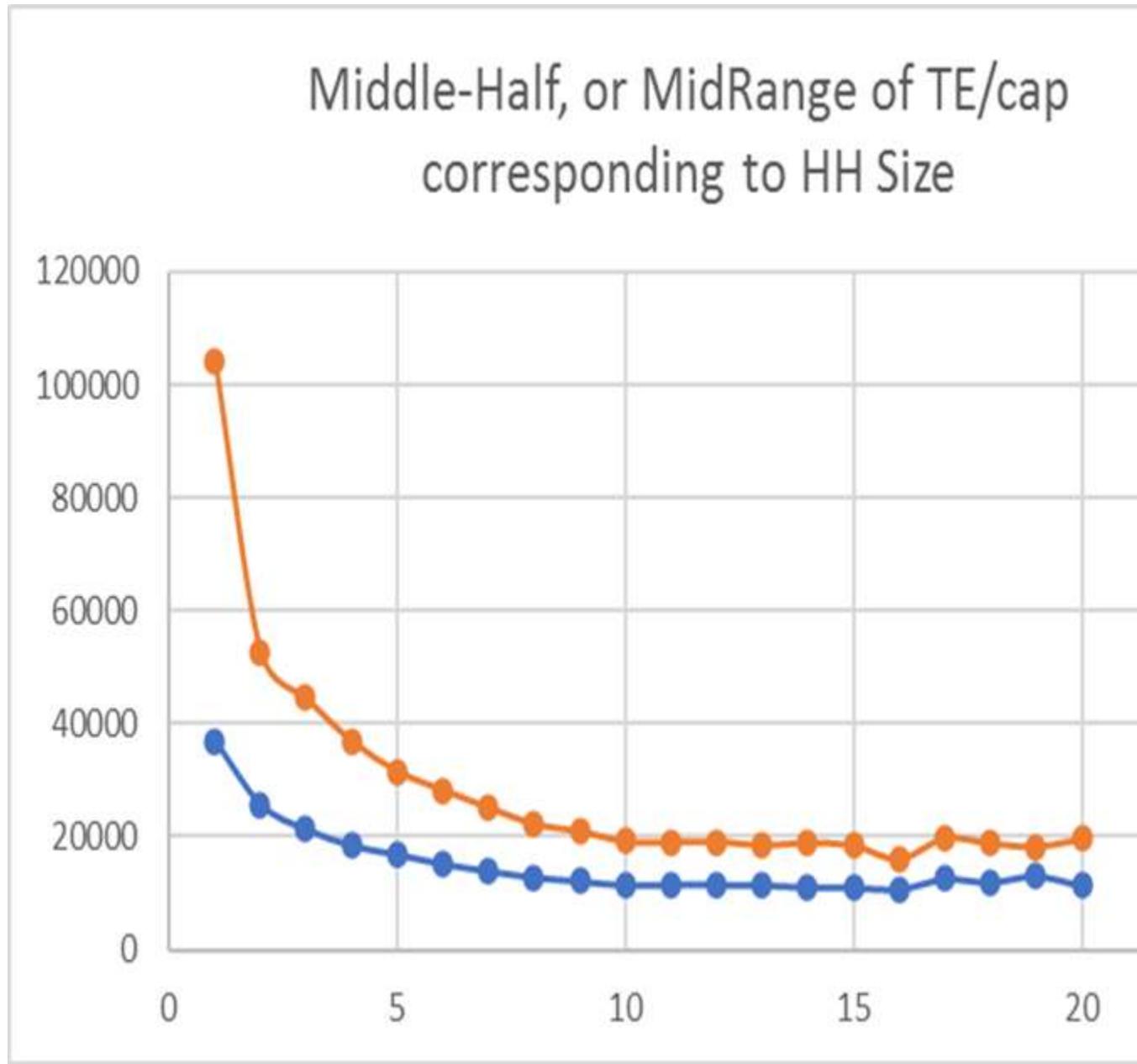
In this table, the second column give the INDEX number which has the LAST entry for the HH Size indicated. That is, HH Sizes are sorted so that the first 158 entries are 1's and the next 649 entries are 2's and so on. The 159 next to the 1 shows that the "1" entries for HH Size END at the cell A159. How do we know this? Because there are 158 HH's of this size, and the entries start at A2, since the cell A1 contains the NAME HH Size, and not data. Similarly, we get the 808 by adding the number 649 of HH's of size 2, to the previous value of 159. This gives us cell A808 as the last cell in the A column which contains the value 2. The entries for HH Size of 3 start at A809. Adding the number of HH's in each category to the previous number gives us the second column of the table above.

The next step is to compute the range of cells for EXP/cap corresponding to each category of HH Size. HH Size of 1 goes from A2 to A159, so the exp/cap of these HH's goes from B2 to B159. HH Size of 2 goes from A809 to A1840, so the corresponding Exp/Cap of this group goes from B809 to B1840. The table above provides the starting and ending entry of the Exp/Cap for each of the groups corresponding to each of the HH Size going from 1 to 20.

The next step is easy. We simply compute the quartiles using the quartile command. For example, Quartile(B809:B1840,1) gives us the first quartile of Exp/Cap for all HH's in HH Size 3. Using this method, we can get all three quartiles for each of 20 groups of HH Sizes going from 1 to 20:

| H H Size | Q1 | Q2 | Q3 | H H Size | Q1 | Q2 | Q3 |
|-------------|-----|-----|------|-------------|-----|-----|-----|
| 1 31 | 370 | 591 | 1042 | 1 94 | 115 | 145 | 190 |
| 2 65 | 257 | 352 | 5272 | 1 60 | 115 | 146 | 191 |
| 3 49 | 213 | 295 | 4453 | 1 62 | 114 | 140 | 185 |
| 4 96 | 184 | 251 | 3694 | 1 88 | 110 | 134 | 189 |
| 5 03 | 169 | 221 | 3158 | 1 34 | 110 | 142 | 184 |
| 6 08 | 152 | 204 | 2826 | 1 57 | 107 | 128 | 160 |
| 7 73 | 139 | 181 | 2514 | 1 38 | 125 | 153 | 197 |
| 8 27 | 128 | 165 | 2228 | 1 74 | 119 | 137 | 188 |
| 9 72 | 121 | 153 | 2098 | 1 55 | 130 | 160 | 182 |
| 0 46 | 114 | 144 | 1920 | 2 67 | 113 | 135 | 195 |
| | 75 | 0 | | 0 | 93 | 89 | |

The Q2 gives the median lines, which we have discussed in the previous lecture. In this lecture, our focus is on the MIDRANGE, which is the set of values from Q1 to Q3. This is where the middle half of the data is, with 25% being below Q1 and 25% being above Q3. This gives us a good summary picture of the distribution. The required plot is easy to make in EXCEL, and comes out as follows:



The lower blue line is Q1, while the upper orange line is Q3. Between the two is half of the data. The midrange measure the variation in the data. We see that there is a LOT of variation in Exp/Cap in the lower HH Sizes. For HH Size 1, the middle half of the population has exp/cap ranging from below 40,000 to above 100,000. But after HH Size 10, the Q1 and Q3 are close to each other, showing small range of variation. Furthermore, the picture does not change much beyond this point. The above graph shows the stochastic relationship between the HH Size and the Exp/Cap. The HH Size does not determine the Exp/Cap which can fluctuate quite widely at each HH Size. But we can look at the distribution and how it changes. The above graphs show how the dispersion or SPREAD of the incomes DECREASES as the HH Size increases, and stabilizes after HH Size reaches 10.

Next, we consider the MIDRANGE of the HH Size for given Exp/Cap. While the HH Size was fixed at integer values, the Exp/Cap is free to vary and is different for every HH. This means we cannot fix it at one number. Instead, we can classify it into several small ranges, and then work with each range separately. Since $20 \times 775 = 15,500$ and our HIES data set is 15,509 points, it is convenient to create 20 groups of exp/cap with size approximately 775. To find the range of the exp/cap groups, we must first sort the data by this variable. Then we simply count the entry after every 775 or 776 cells to find the income ranges for each of the 20 groups. This leads us to the following table:

| Gro ups | Lo w exp/cap | Hig h exp/cap | Gro up Size | Gro ups | Gro ups | Lo w exp/cap | High exp/cap | Gro up Size |
|------------|-----------------|------------------|----------------|------------|------------|-----------------|-----------------|----------------|
| 1 | 196 6 | 960 9 | 775 | 11 | 194 55 | 2081 9 | 775 | |
| 2 | 961 4 | 112 20 | 776 | 12 | 208 21 | 2230 8 | 776 | |
| 3 | 112 21 | 123 24 | 775 | 13 | 223 09 | 2401 8 | 775 | |
| 4 | 123 24 | 133 59 | 776 | 14 | 240 24 | 2615 1 | 776 | |
| 5 | 133 59 | 142 75 | 775 | 15 | 261 54 | 2864 8 | 775 | |
| 6 | 142 75 | 152 35 | 776 | 16 | 286 51 | 3190 9 | 776 | |
| 7 | 152 36 | 161 99 | 775 | 17 | 319 16 | 3709 8 | 775 | |
| 8 | 162 00 | 172 35 | 776 | 18 | 371 00 | 4556 5 | 776 | |
| 9 | 172 36 | 182 72 | 775 | 19 | 455 76 | 6429 7 | 775 | |
| 10 | 182 73 | 194 54 | 776 | 20 | 643 17 | 1268 708 | 775 | |

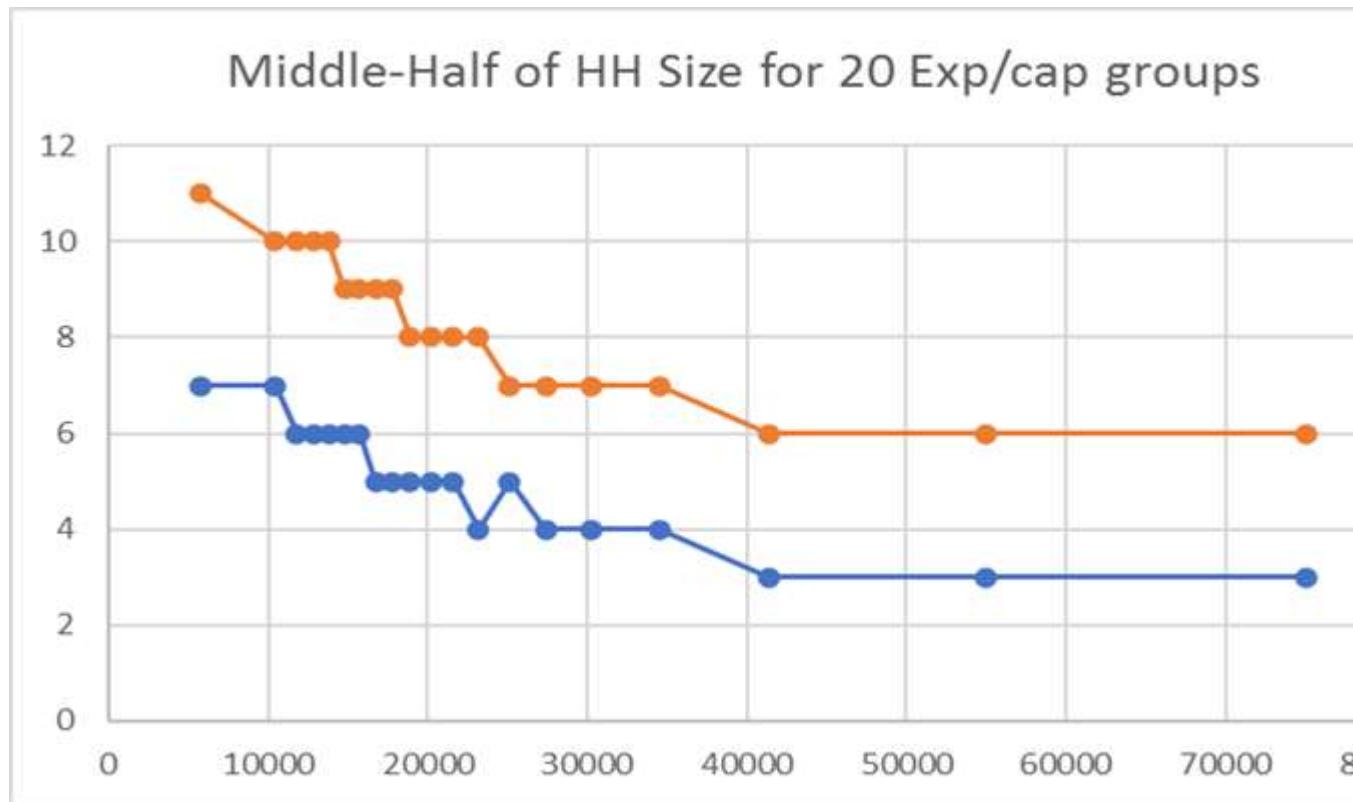
This gives us the 20 groups where exp/cap ranges from the low to high value as specified in table above. This creates 20 income groups, each having size 775 or 776. For group size of 775 each we get a total of 15500 points leaving 9 extra points. By adding one data point to every other group, we can accommodate these 9 extra points of data as we have done above. After have fixed the income ranges, we need to compute the quartiles for the HH Size in each of the 20

income groups. This involves first calculating the range for each group and then applying the quartile command to that range. This leads to the following table:

| Lo | Hi | q1 | q2 | q3 |
|--------|--------|----|----|----|
| A2 | A776 | 7 | 9 | 11 |
| A777 | A1552 | 7 | 8 | 10 |
| A1553 | A2327 | 6 | 8 | 10 |
| A2328 | A3103 | 6 | 8 | 10 |
| A3104 | A3878 | 6 | 8 | 10 |
| A3879 | A4654 | 6 | 7 | 9 |
| A4655 | A5429 | 6 | 7 | 9 |
| A5430 | A6205 | 5 | 7 | 9 |
| A6206 | A6980 | 5 | 7 | 9 |
| A6981 | A7756 | 5 | 6 | 8 |
| A7757 | A8531 | 5 | 6 | 8 |
| A8532 | A9307 | 5 | 6 | 8 |
| A9308 | A10082 | 4 | 6 | 8 |
| A10083 | A10858 | 5 | 6 | 7 |
| A10859 | A11633 | 4 | 5 | 7 |
| A11634 | A12409 | 4 | 5 | 7 |
| A12410 | A13184 | 4 | 5 | 7 |
| A13185 | A13960 | 3 | 5 | 6 |
| A13961 | A14735 | 3 | 5 | 6 |
| A14736 | A15510 | 3 | 4 | 6 |

For each of the income group, the first two columns give the low and high index for the corresponding HH Sizes in that group. Then we apply the QUARTILE command to the entries in the first two columns. We have already plotted Q2 or the median in the previous portion of this

lecture. We now provide the midrange plot of the middle half of the HH Size in each group separately. This is graphed in the following picture.

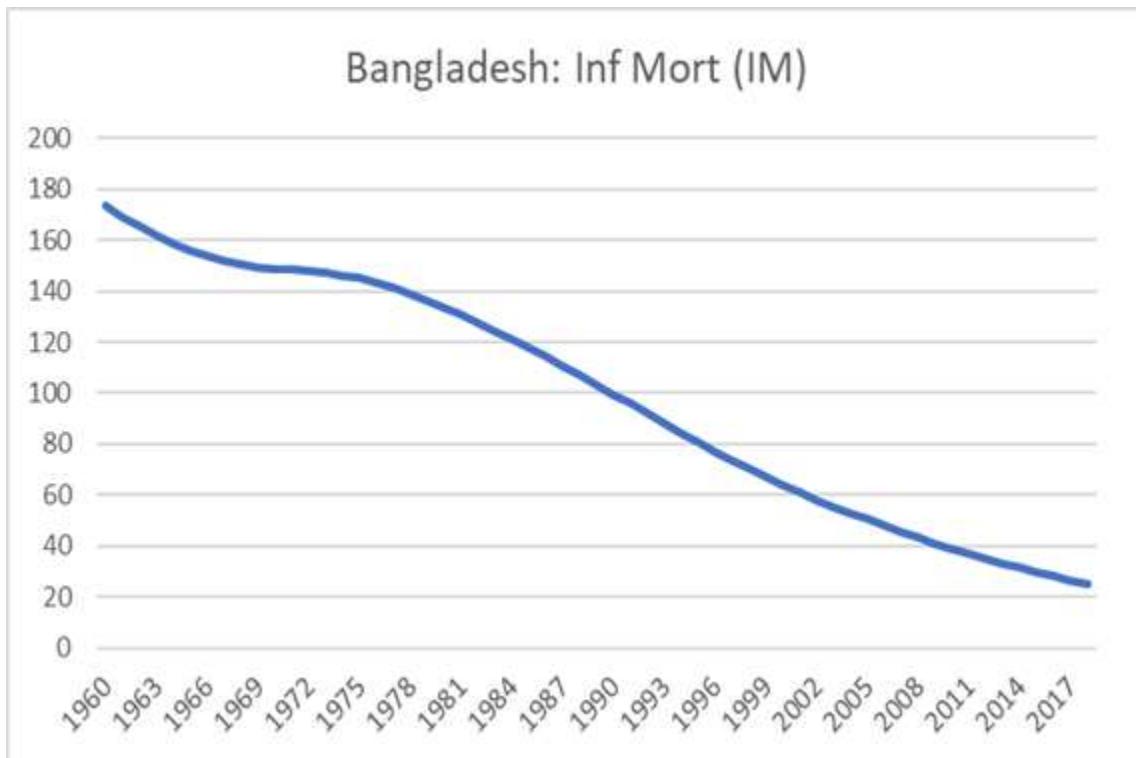


This gives us a picture of the distribution of the HH Size as it varies by exp/cap. In the lowest expenditure groups, the midrange goes from HHSIZE=7 to HH SIZE= 11. This means that about 50% of the HH's in the lowest expenditure category have HH Sizes ranging from 7 to 11, while 25% of them have HH Size below 7 and another 25% have HH Size above 11. This MIDRANGE keeps shifting down as the exp/cap increases, and stabilizes at the range [3,6] for the top 3 income groups. This means that among wealthy families, about 50% have HH Size between 3 and 6 – corresponding to having between 1 and 4 children. Compared to the richer countries, these numbers are on the high side.

The purpose of this lecture was to explain the concept of a Stochastic Relationship: a relationship between VALUE of X and DISTRIBUTION of Y. In the previous lecture, we studies the Median-Line: Half of the data points are ABOVE median line and half are below – gives a picture of the CENTER of the distribution or the second quartile Q2. In this lecture we focus on the Mid-Range of the Data, which lies between Q1 and Q3. The Middle Half of the population is BETWEEN these two quartiles, with 25% is below Q1 and 25% above Q3. The Mid-Range provides a picture of the SPREAD of the data. A more technical term is the Interquartile Range, which is the difference between Q3 and Q1. This provides a numerical measure of the spread, while the MIDRANGE plot provides a VISUAL description of this spread.

6D Evolution of Distribution of Infant Mortality (IM) across time

We want to study how the “distribution” of Infant Mortality (IM) changed over time. This is one type of a stochastic relationship: we study the distribution of IM as a function of time. It is important to clearly differentiate between the distribution of IM from the IM for any country or group of countries. For example, we can study how IM in Bangladesh changed over time. The IM number for each for Bangladesh can be plotted as follows:

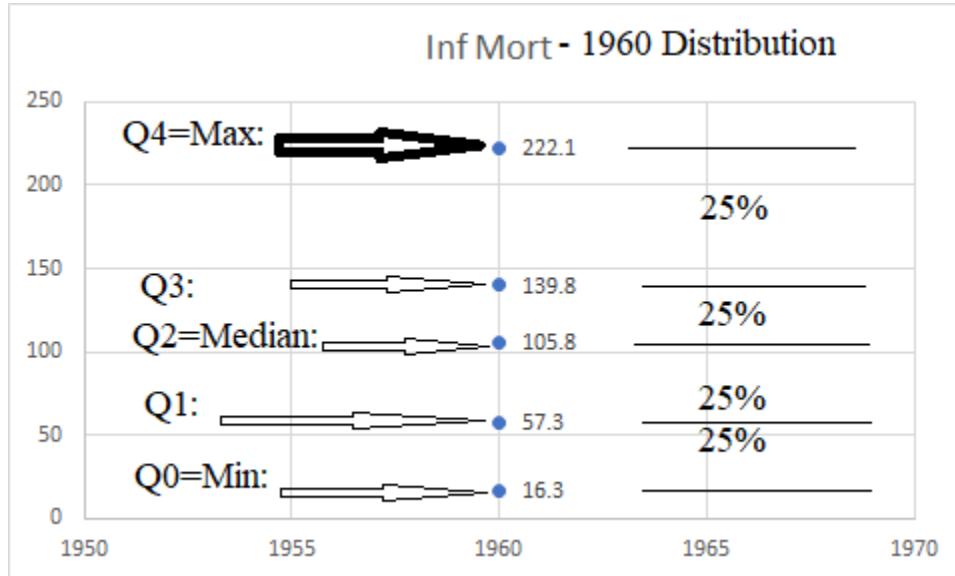


Bangladesh IM starts at 173 and declines to 25 from 1960 to 2017. Similarly, we can draw the graph of IM, and how it changes over time for any country. BUT, we want to look at the DISTRIBUTION of IM, and how it changes over time. This does not come from individual countries, but from the WHOLE collection of countries. This is an example of a Stochastic Relationship: How does DISTRIBUTION of IM change over time?

We must clearly differentiate between DISTRIBUTION and individual country. No ONE country, or group of countries, gives us the Distribution. The distribution comes from looking at IM in ALL the countries, and does not translate to any single country. We can't go from individual cases to distribution, and we cannot go from distribution to individual cases.

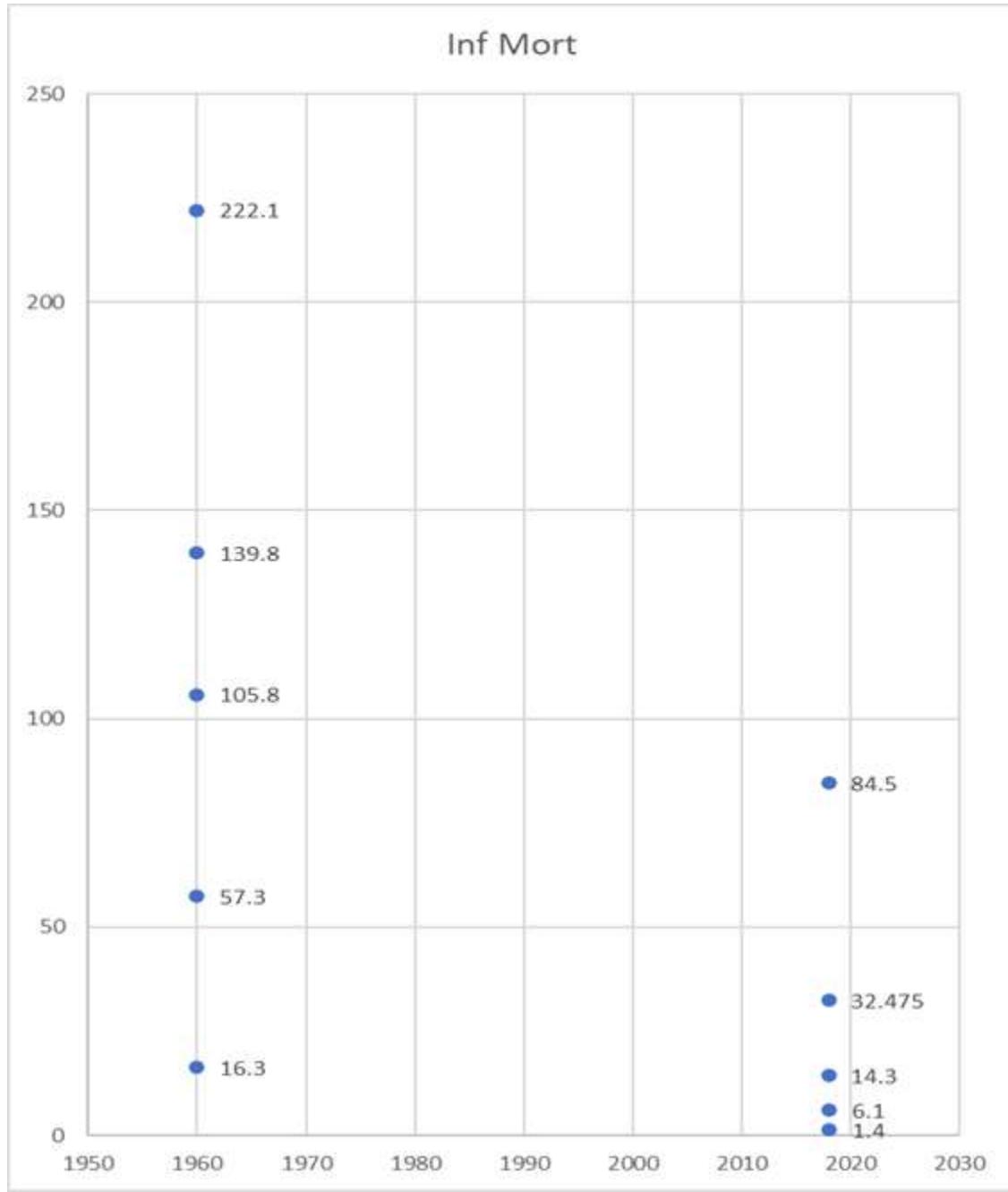
The technical definition of the distribution is a bit complicated, and will be explained in detail later. A visual description of the distribution is the Histogram, which have studied earlier. Because the Histogram is a complex entity, studying how it changes over time creates difficulties. Instead, we will look at the five-point SUMMARY of distribution, based on the five quartiles – Q0, Q1, Q2, Q3, Q4. These five numbers create four intervals such that 25% of the

countries belong to each of them. We illustrate this by looking at the Distribution of Infant Mortality in 1960:



The Five Quartiles provide a first approximation PICTURE of the distribution. Between each two quartiles, we have 25% of the countries. 105.8 is MEDIAN – half countries are below and half are above. 25% have IM between 16.3 and 57.3 – these are the BEST 25% countries, with the lowest IM's in 1960. To understand the difference between distribution and individual countries, note that the Distribution tells us how the countries are spread out over the range going from Minimum to Maximum. But it DOES NOT identify the countries: which one falls where? For example, know the range of IM's for the best 25% of the countries, but the distribution does tell us WHICH countries belong to this category. We cannot find out WHICH country has the best IM=16.3 and which one has the worst IM=222.1. We can of course go back to the data set to find out the answers to these questions, but the distribution does not contain this information.

Similarly, individual country data DOES NOT give us the distribution. To learn distribution, we must know about ALL the countries, and we must ARRANGE the data to display the distribution. Since we are interested in the changes in distribution, let us do a comparison the (five-point summaries) of the distribution of IM's from 1960 to 2018. This is plotted below:

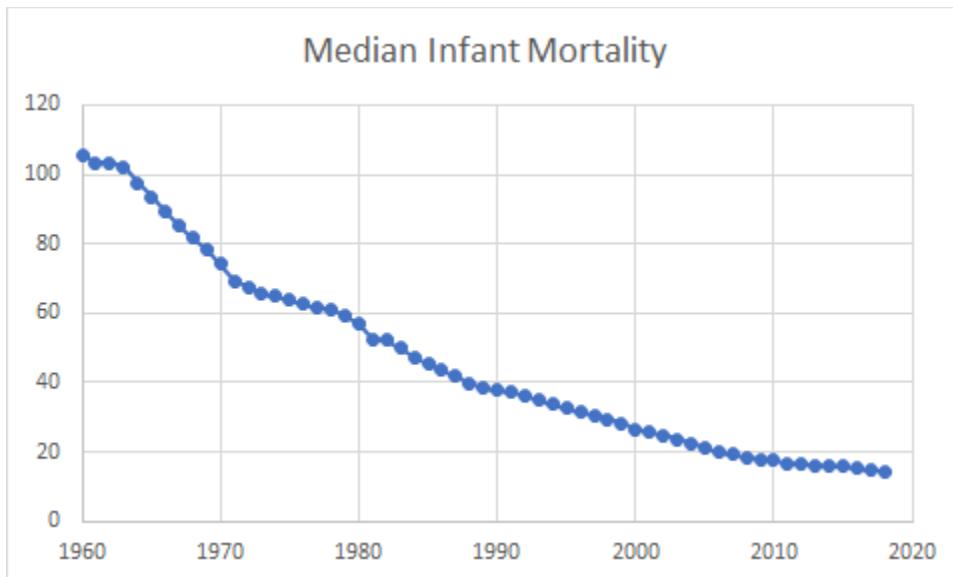


We can see the dramatic change in the distribution from 1960 to 2018. In 1960, the best 25% of the countries had IM in the range 16.3 to 57.8. This changed to 1.4 to 6.1 in 2018. Similarly, the worst Quartile has IM's in the range 139.4 to 227.1. In 2018 this changed to: 32.5 to 84.5. Both the best and worst group of countries had dramatic improvements in IM. This shows that there was comprehensive change everywhere. But the data does not tell us about “What are the causes of this change?”

Before proceeding with our analysis, we come back to the Fundamental Questions of Real Statistics:

1. **Q:** Why are we looking at these Infant Mortality numbers? **A:** Death of Children is an extremely significant life event, it matters.
2. **Q:** What do the numbers mean? **A:** Number of deaths within one year from 1000 Live Births.
3. **Q:** How are they computed? **A:** Varies with country – surveys, random samples, etc. In countries where all births are registered, it is easy to cross-match records to find out how many children die within one year, from a randomly chosen group of 1000 births. In countries like Pakistan, where most births are not recorded, it tends to be more difficult.
4. **Q:** What will we do with the analysis? **A:** Understand causes of high IM and try to remove them.

Coming back to our main question: How did distribution of IM change across time? This can be captured by looking at the five quartiles of the data, and how they change across time. We start by looking at the Median IM among all countries in WDI, as plotted below:



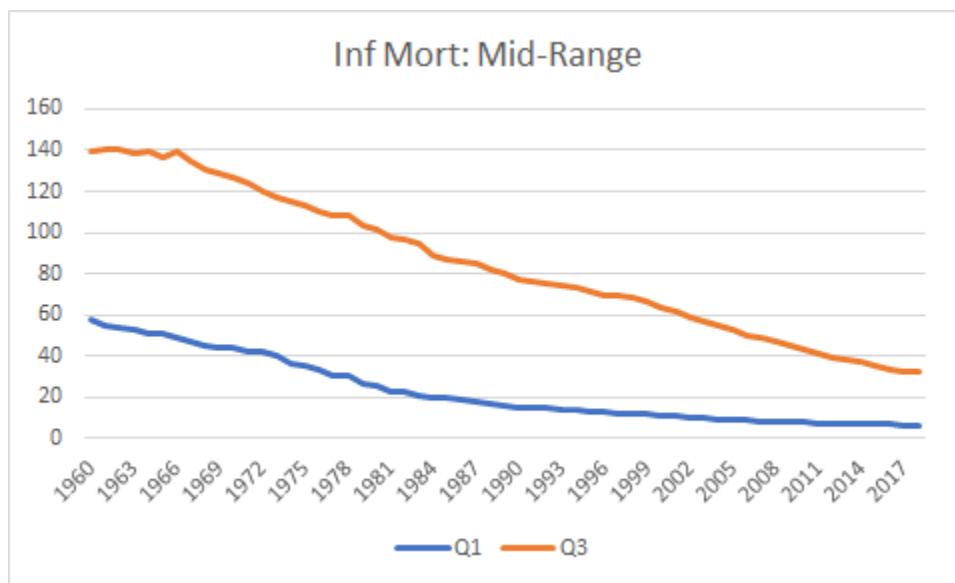
The graph shows that it has declined, smoothly, from 106 to 14! A HUGE reduction. Note that Half (50%) of the all the countries in the WDI data set now have IM's below 14. Half of them are above this fairly low number. WHY has IM declined across the board? The numbers have no answer. TO learn about this, we need to look at CAUSES of IM and see how deaths due to different causes have changed over time. This will allow us deeper insight into the causes of this rapid decline. It is well-known that One important factor is MORE in-hospital births. But how much this matters will be considered later.

From the statistical perspective, the Median is the CENTER of the distribution. It can be considered as the best ONE NUMBER summary of the distribution. Standard Statistics textbooks and software use MEAN as CENTER and best one number summary. This is based on the ASSUMPTION that the data is NORMAL. Such an assumption is rarely justified on real data sets. The MEAN does not have any direct and intuitive interpretation in general (non-normal)

data sets. In contrast, the Median tells us the HALF of the countries are above, and half are below. This is easily understood intuitively. 50% of the countries have IM at 14 per 1000 or below. This cannot be too hard to achieve.

What can we hope to learn by studying these curves, which tell us about the changes in the distributions of IM over time? On many occasions in the past, this macro-level knowledge of the big picture has led to Micro-level Investigations of some importance. For example, doctors were puzzled by a sudden rise in a certain rare type of Lung Cancer. Studying the data, they realized that this coincided with a rise in the use of Cigarettes. Then, there was a debate and discussion about the causal relationship, where tobacco companies fought to deny the link, while many sought to establish it. Today, a similar battle is being fought about the massive increase in the use of Sugar, and the links Diabetes, Heart Attacks, and many other diseases. The sugar industry is denying the links, while the empirical evidence is mounting against the overuse of sugar. The point here is that studying big trends CAN lead us to knowledge about real world mechanisms.

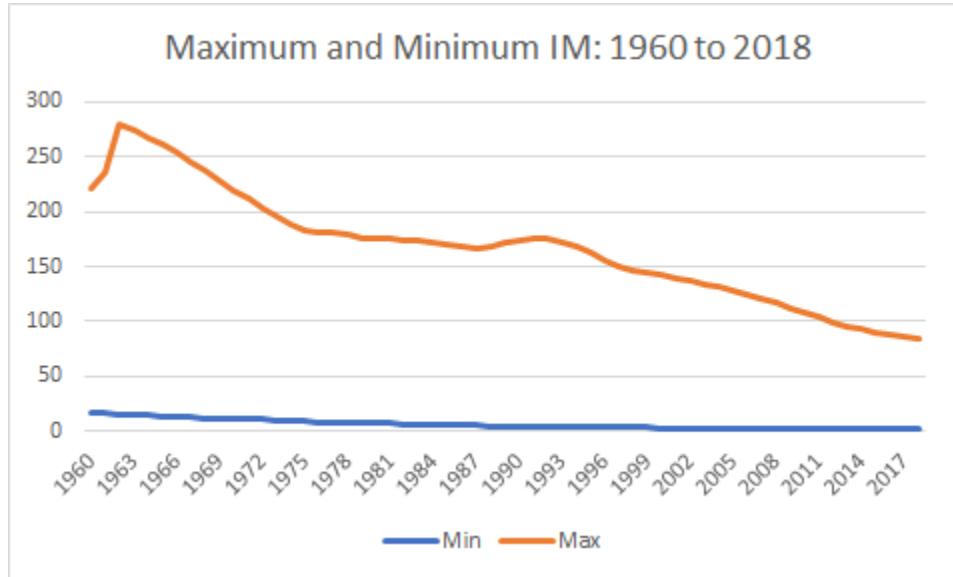
The median is just one SMALL part of the DISTRIBUTION which we want to study. The five quartiles provide us with a more complete picture. Next, we look at Q1 and Q3, the first and third quartiles. The range between Q1 and Q3, we call the Mid-Range, as it covers the Middle Half of the DATA. For the Infant Mortality data, this can be pictured as follows:



Half of the countries lie within the range covered by Q1 and Q3, blue and yellow lines. We can see that the DISPERSION, or SPREAD, of the distribution has decreased over time. In 1960 the mid-range was (57,140), while in 2018 it was reduced to (6,33). This is a massive reduction. Such a big change must have causes. We need to examine the real world context, in order to search for causes. Note that the graph tells us that 25% have IM < 6 ! This is impressive, it means that less than 6 children die within one year, from 1000 live childbirths. Since a large number of countries have achieved this goal, it should be possible to replicate this achievement

in other countries by a careful study of the causes of this improvement, and the application of this knowledge.

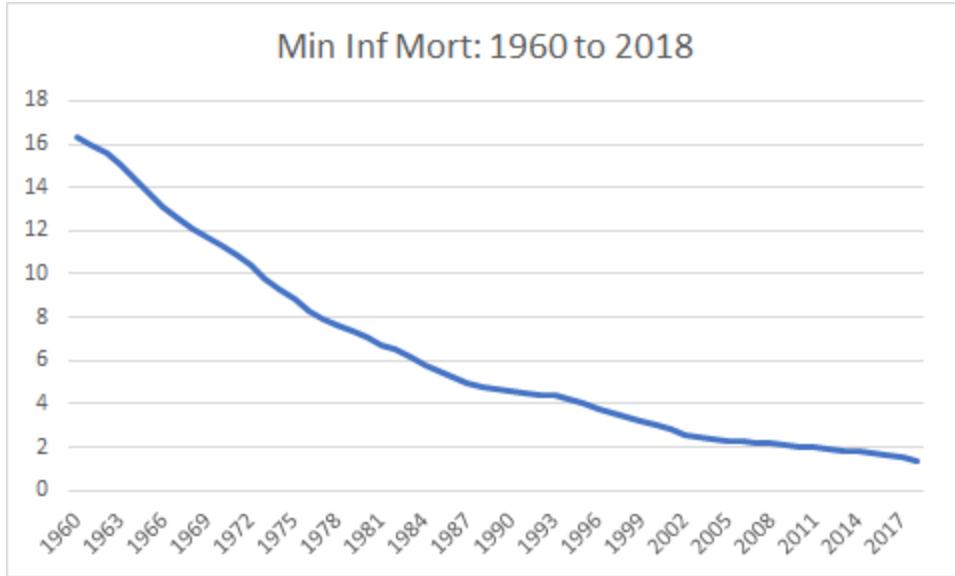
Next we will look at the Range: from the Minimum (Q0) to Maximum (Q4). This is pictured below:



With occasional exceptions, the maximum has been going down by a lot. This means that the worst-case country has gone from over 250 to below 100. Again, this tell us that the reduction in IM is world-wide, over all countries. Why do we have a few INCREASES in the maximum, as the graph shows? This is because we are looking the worst-case. This may have to do with WAR or special events in ONE country. The maximum is erratic and does not show the general trend of all countries.

From the graph, note that the Range has SHRUNK substantially. Where the difference was more than 250 between the best and worst, this difference is now only 80 or so.

It is worth noting that because the Maximum is so large, variations in Minimum are not visible. The SCALE of graphs is VERY IMPORTANT. When we scale IM from 0 to 300 then small changes become invisible. This will become clear when we plot the minimum (Q0) separately. Here is a graph of the BEST Infant Mortality: Minimum



When the vertical axis goes from 0 to 300, a change from 15 to 1 is barely noticeable. On a separate graph, we see clearly the strong declining trend in the minimum. As we can see, the Minimum IM went from 16 to 1.4. We see that with good medical care, nearly ALL children born can survive to 1-yr. We need to study the countries with the best Infant Mortality rates, to see the lessons we can learn about how to reduce infant mortality. Learning about “causes” is important, but these causes are not present in the data. Rather, learning about causes, requires SHOE LEATHER. That is, we need to go out into the real world and study it. HOW can we learn more about causes?

There is the External Approach: More Data on timing – deaths within one week, one month, and one year have different causes. Early deaths due to no pre-natal monitoring and micro-nutrients, as well as basic mismanagement of birth, and early complications. The 80-20 rule applies. 80% of early deaths (mother & child) are easily preventable. No fancy technology required. The internal approach requires a deeper exploration of the unobservables and the unquantifiable. For example, a lot depends on Social Norms. In some cultures, home-births are preferred to hospitals. In such cases, we need to create changes in SOCIAL NORMS. We need to create social awareness of the problem and to prioritize a solution of the problem in order to create change. None of this requires a lot of money. The example of CUBA is amazing. They have built an excellent health system, with minimal resources. Thus wealth is not required to lower Infant Mortality substantially.

Some concluding remarks on the PROCESS of acquisition of knowledge through numbers. Data is collected on some types of number but not on others – WHY? This is because our theoretical pre-conceptions GUIDE collection of data, both what is collected and what is NOT collected. Data that our theories tell us are important will get collected, while other types will not be. Unbiased study of data, combined with knowledge of the real world, leads to QUESTIONS. Tentative answers to these questions are HYPOTHESES. To test hypotheses, we need to collect more data. For example, in-hospital births have contributed (how much?) to reduction in IM. To answer this, we need data on proportion of in-hospital births vs at-home.

From this increase, we can calculate if this is sufficient to explain the decline in mortality. It turns out this helps, but is not sufficient to explain the dramatic decline in IM. Therefore, we need to look for other causes. The point is that theories guide us to collect data. Then the data may not be fully compatible with the theory and may indicate other avenues (hypothesis) of interest. Examining the data to find interesting patterns which reveal something about the real world is the art of statistics. Once we come up with some new hypotheses about how the world works, we will need to collect more data to confirm. This back-and-forth process between data and hypotheses is how knowledge advances.

6E Comparing Progress of Countries on Infant Mortality

In the past few lectures, we have looked at the data on Infant Mortality. In this lecture, we will address the question of “How have countries performed, in terms of reducing IM?”. As discussed in an earlier lecture, rankings cannot be done objectively. For a subjective ranking, we need to consider the PURPOSE for which the ranking is being done. We now consider this topic.

What is the GOAL of evaluating performance of different countries with respect to Infant Mortality? There is an important goal: Find out the winners, and study their ways, to learn about good strategies for lowering IM. We can also study the losers, and the causes for their failures, in order to avoid them. Note that this will involve going beyond the data to study health systems in the high ranking countries to learn lessons about how to reduce Infant Mortality. This leads us to the META-Principle: Data Analysis provides CLUES about what is happening in the real world. It is NEVER complete in itself. CLUES must be followed up by study of real world to see if the clue points to some real OR NOT. This principle is directly opposed to current statistical practice, which is limited to the analysis of numbers by themselves. One important consequence of this meta-principle is that it becomes important to distinguish between valid clues and false and misleading ones. Sometime this can be done by looking at the numbers, but at times, it would be necessary to look at the real world to see if the clue points to something which exists.

We come back to the main question: “How to rank countries in terms of progress in lowering IM?”. One simple possibility: Compare IM 1960 with IM 2018. How many MORE lives of children are saved over the 58 year span that the WDI data covers? One problem with this strategy is that 1960 statistics are not available for many countries. A solution to this problem is to use the Maximum IM over all the years as the worst IM, and compare to IM 2018, what the country has achieved in the most recent year for which data is available. According to this strategy, we calculated the Maximum IM for each country (IM Max) and subtract IM 2018. This gives us a Change (IM Max – IM 2018) score for each country. Now we rank all the countries by this score. This leads us to the following table of countries with top scores:

| Country | Change | Max | 2018 |
|----------|--------|-------|------|
| Yemen | 236.5 | 279.4 | 42.9 |
| Oman | 206.5 | 216.3 | 9.8 |
| Maldives | 202.7 | 210.1 | 7.4 |

| | | | |
|-------------|-------|-------|------|
| Egypt | 192.8 | 210.9 | 18.1 |
| Nepal | 189.4 | 216.1 | 26.7 |
| Afghanistan | 188.6 | 236.5 | 47.9 |
| Malawi | 175.7 | 211 | 35.3 |
| Tunisia | 167.5 | 182.1 | 14.6 |
| Turkey | 163.2 | 172.3 | 9.1 |
| Bhutan | 162.6 | 187.4 | 24.8 |
| Liberia | 157.6 | 211.1 | 53.5 |
| Cambodia | 153.7 | 177.7 | 24 |

This ranking leads to Yemen as top ranked. This is a bit odd, since Yemen is war-torn country and must have a very poor IM rate currently. We also note that nearly all of the countries in this list are those which have terribly high Maximum IM. This leads to a concern that perhaps this method is NOT identifying the countries which have done well. The ABSOLUTE change in IM may not be a good measure of progress. Suppose that a country was at IM=500 and reduced to IM=250, it would be ranked as best, because the change is very high. This would be true even though this country would be the WORST in terms of IM at all times. Something like this actually happens in this data set. The worst IM is Q5(=Max IM) goes from 222 to 85. The change is 137 which is very large, making it a good country under this criterion. The problem is that countries which are very bad in terms of IM have the largest room for improvement.

How can we improve this criterion? One Solution: Switch to PROPORTIONAL change. If the worst country goes from 222 to 85, this is a proportional reduction of $137/222=61.7\%$. If another country goes from 105 to 1, this is a proportional reduction of $104/105=99\%$. So the proportional reduction is larger, even though the absolute reduction is not. We can look at the proportional reduction by looking at Proportion = $\text{LOG}([\text{IM Max}]/[\text{IM 2018}])$. Using this criterion produces the following rankings, where we list the top countries:

| Country Name | Proportion | Max | 2018 |
|--------------|------------|-------|------|
| S Korea | 1.470 | 79.7 | 2.7 |
| Maldives | 1.453 | 210.1 | 7.4 |
| Portugal | 1.434 | 84.2 | 3.1 |
| Oman | 1.344 | 216.3 | 9.8 |
| Bahrain | 1.341 | 133.7 | 6.1 |

Author Last Name/Book Title

| | | | |
|--------------|-------|-------|------|
| Chile | 1.325 | 131 | 6.2 |
| U A E | 1.315 | 134.1 | 6.5 |
| Turkey | 1.277 | 172.3 | 9.1 |
| Saudi Arabia | 1.261 | 109.4 | 6 |
| Italy | 1.230 | 44.2 | 2.6 |
| Japan | 1.228 | 30.4 | 1.8 |
| Libya | 1.205 | 163.4 | 10.2 |
| Finland | 1.196 | 22 | 1.4 |
| Singapore | 1.187 | 35.4 | 2.3 |

However, in this lecture, we will consider a THIRD method for evaluating improvements. We are interesting in knowing HOW MANY COUNTRIES were overtaken by a given country. If country goes from MEDIAN to MINIMUM it overtakes 50% of the countries – this is a BIG achievement. So the solution we will focus on in this lecture involves a: Switch to RANK. Instead of looking at the number IM, replace IM by RANKING of the country. There are 192 countries on which data is available. The worst country has rank 192 and the best has rank 1. We would like to see how the RANK of the country changes over time. Difference in RANK captures RELATIVE improvement, measures number of countries overtaken directly. That is, it tells us about how many countries have been left behind by the given country, in the race to lower IM. The next table gives us the effect of a Switch to Ranks from our initial criterion of the Absolute difference in the number of IMs.

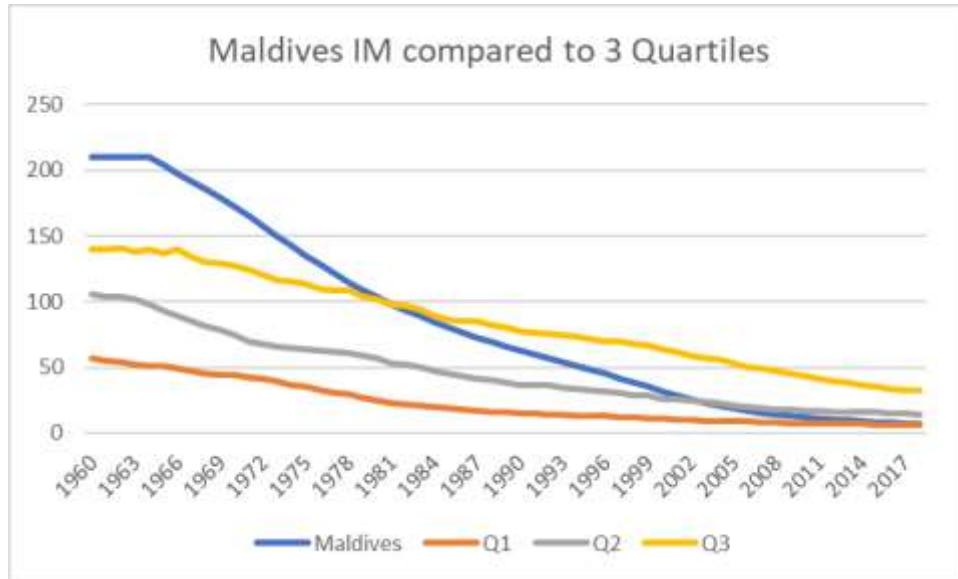
| Country | Code | D IM | D x IM | Ma 2018 | IM k Max | Ran k 2018 | Ran k 2018 | Diff Rank |
|----------|------|---------|-----------|------------|-------------|---------------|---------------|--------------|
| Maldives | MDV | 202.7 | 210.1 | 7.4 | 184 | 60 | 4 | 12 |
| Oman | OMN | 206.5 | 216.3 | 9.8 | 190 | 72 | 8 | 11 |
| Turkey | TUR | 163.2 | 172.3 | 9.1 | 170 | 71 | | 99 |
| Libya | LBY | 153.2 | 163.4 | 10.2 | 168 | 75 | | 93 |
| Bahrain | BHR | 127.6 | 133.7 | 6.1 | 141 | 49 | | 92 |

| | | | | | | | | |
|---------|---------|---|-----------|-----------|-----|-----|-----|----|
| U A E | AR E | 6 | 127. 1 | 134. | 6.5 | 142 | 57 | 85 |
| Chile | CH L | 8 | 124. | 131 | 6.2 | 138 | 53 | 85 |
| Tunisia | TU N | 5 | 167. 1 | 182. 6 | 14. | 176 | 98 | 78 |
| Korea | KO R | | 77 | 79.7 | 2.7 | 93 | 17 | 76 |
| Egypt | EG Y | 8 | 192. 9 | 210. 1 | 18. | 185 | 112 | 73 |

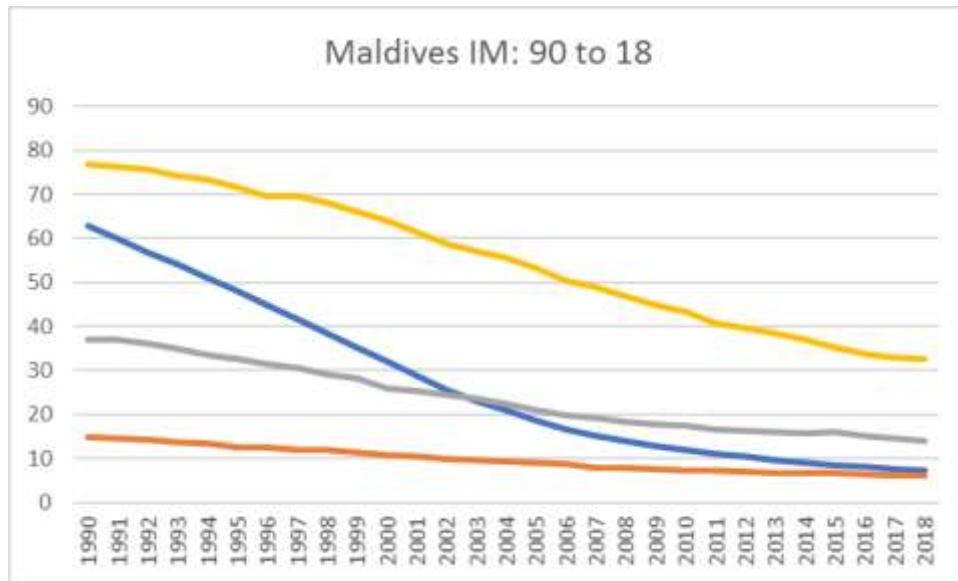
With this change, Maldives becomes the top ranked country. The Max IM for MDV was 210.1 which was at rank 184, near the bottom rank of 193. In 2018 MDV was ranked at 60th, which is an improvement of 124 positions. That is the largest change in the WDI data set. As opposed to this, Yemen went from Max IM 279 to IM 2018 = 43. This was at bottom rank 193 with the highest Max IM. In 2018, it was at rank 165 with IM of 42.9. The improvement in rank was only 28, even though the absolute improvement in numbers was very large.

Despite this difference, we note that ALL countries in top positions are among WORST ranking in IM, except for Korea. Many of the countries which came out in the top 10 in the Absolute Difference in IM category are also high ranking with respect to this difference in ranks criterion. The top 3 countries according to Change in RANK criterion are Maldives, which has rank 184 in Max IM, and goes to rank 60 in IM 2018. Similarly, Oman goes from rank 190 to 72, and Turkey from rank 170 to 71. We would like to learn HOW did these countries improve an extremely bad IM figure to a decent one. We note that the Worst ranking CAN improve most in ABSOLUTE terms. A country with rank 190 has the potential to improve by a 100 positions, while a country with rank 20 can only improve by 19 positions. This means that countries which start out with a good ranking DON'T have much chance to appear at top with respect to this criterion. But this is good for our purposes. We want to know how countries with High IM can get to Low IM. Countries which are already very low IM will not be able to offer much in the way of lessons for this purpose. Note that the Proportion Criterion, discussed earlier, gives a better chance to the countries which are already very good in IM to qualify among the top countries. In fact, Korea comes out at top in the proportion ranking and Japan also comes within the top 10.

In the rest of this lecture, we will use graphs to look in greater detail at each of the top 5 countries with respect to difference in rank, to see how their IM changed over time. We start with the top country, Maldives, which appears among the top 10 in all three ways of ranking progress. Maldives went from having near the worst possible Maximum IM to the 1st quartile. This is shown on the following graph:



This shows the impressive progress of Maldives, going from the worst possible IM to near the 25th Percentile, overtaking 75% of the countries in the world in this process. Because the initial IM is so large, the scale of the graph gets squeezed towards the end, and we cannot see what is happening around 2018 very clearly. To overcome this problem, we draw another graph which starts from 1990. This graph shows the gradual decrease to nearly the 25th percentile

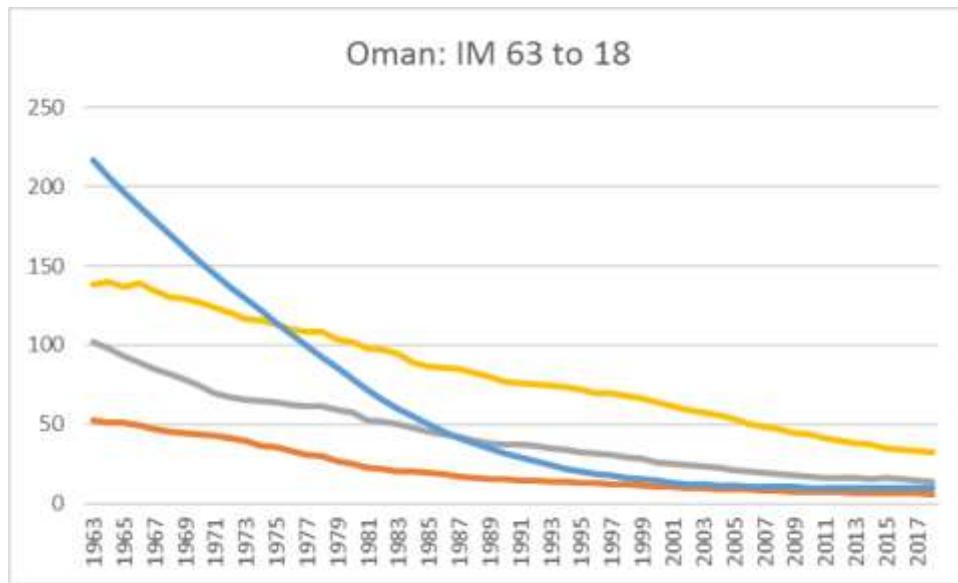


So what are the Lessons we can learn from this graph of Maldives?

1. Maldives is NOT a rich country, but it achieved tremendous reduction in IM. It would be very useful to look carefully at what health systems strategy they used, since this should be replicable by other countries which are not very rich.
2. The graph shows smooth and continuous progress. It shows sustained effort over a long period, and not particular incidents which had a sharp effect on IM either up or down.

3. Maldives reaches 25% by 2018, but the trajectory is flat. This suggests that getting to 25% is relatively easier, further progress is more difficult.
4. This is an Illustration of Pareto principle, or 80-20 principle. The first 80% of the job is relatively easy and takes only 20% of the total effort. Achieving the last 20% is hard, and takes 80% of the effort.

Oman is similar to Maldives in that they have nearly the same Max IM and nearly the same end performance. But Oman is a small oil-rich country, and shows a different trajectory of performance, as in the graph below:

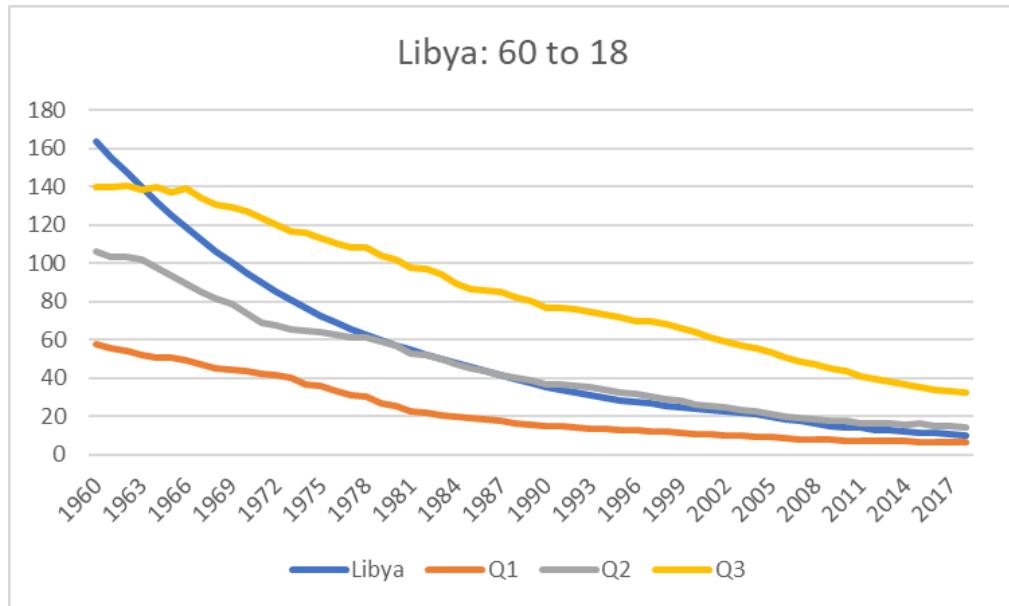


The effect of the oil wealth shows as Oman has a steeper trajectory – more rapid reduction in IM in the early period, so that it achieves 25th Percentile around 1998, almost twenty years ahead of Maldives. On the other hand, after this, there is no further progress, as Oman remains more or less on the 25%, neither improving nor becoming worse. It is clear the oil-wealth permitted Oman to achieve this good performance. However, none of the other oil-rich countries are in the top 10. It is not just having the money, but knowing how to spend it on health systems which matters, and there are lessons to be learnt from Oman in this connection. Also, it is worth noting that Oman did not improve further, beyond the 25th Percentile. So, what are the obstacles to achieving even better performance, even if money is available? Again, this seems to illustrate the 80-20 principle. Getting to the 25th percentile is relatively easy, while improving further is hard. We might be able to learn more about this from a study of Oman, since Oman has the wealth required to improve, but has not been able to do so.

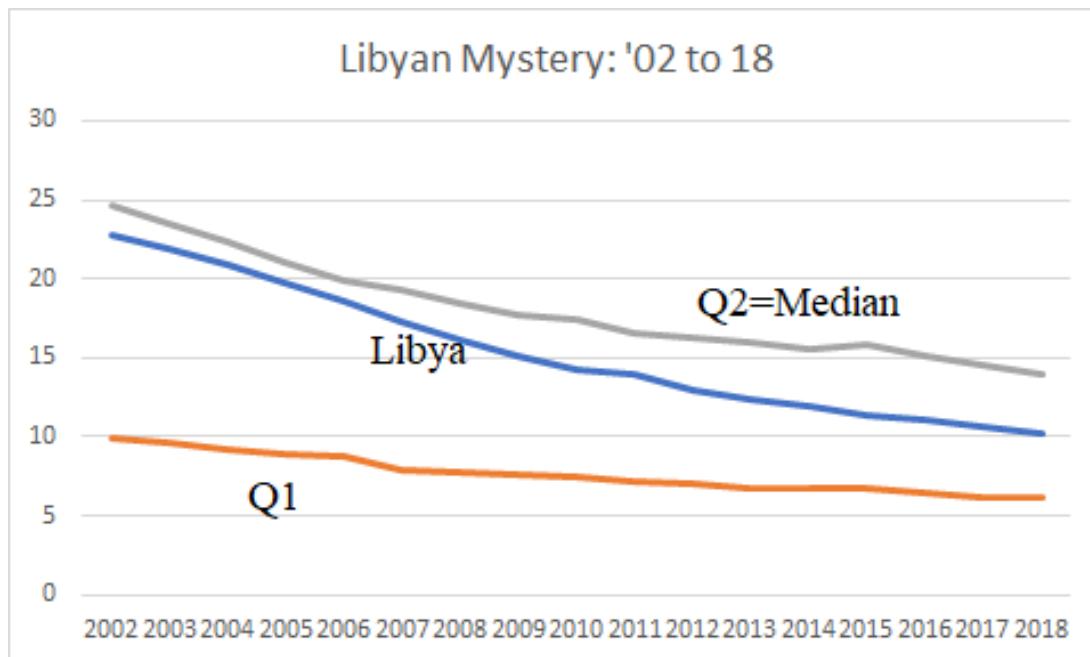
The 3rd and 4th countries are Turkey and Libya, which have similarly start and end points for IM but rather different trajectories. These are shown in the graphics below:

Turkey progresses smoothly from around IM=170 in 1960 to the 25th percentile in 2018, overtaking more than 50% of the countries in this process. There must be lessons to be learnt

from Turkey, since they have made much more progress than other countries. Libya shows a different trajectory:

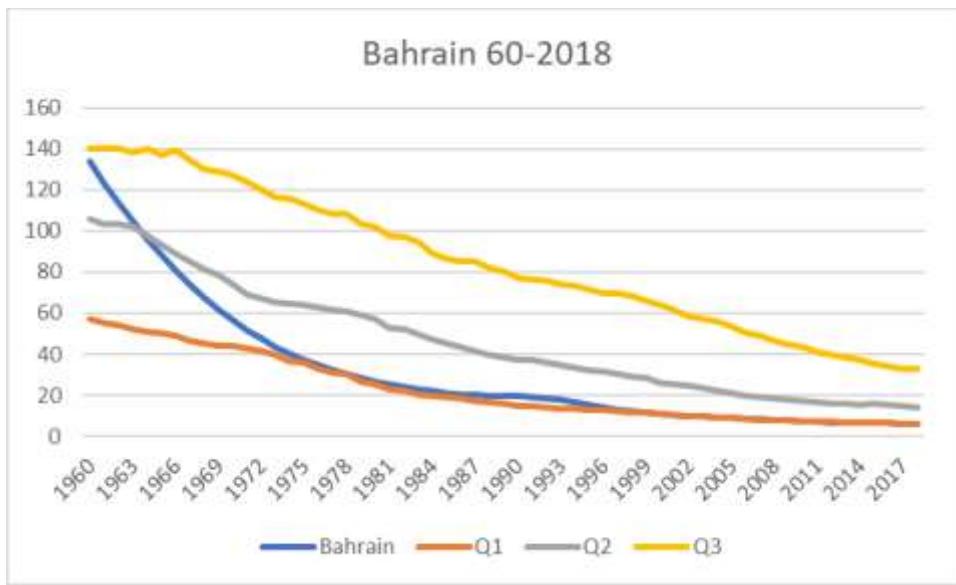


Libya made much more rapid improvements in IM, arriving at the median in 1978, while Turkey arrived at median IMR much later in 2002. Libya's progress shows the effects of good use of oil wealth. Libya had in place comprehensive social welfare systems, with free healthcare and education for all of the population. This was certainly one of the causes of the rapid reduction in IM. But the curve plotted above is surprising. The Libyan Mystery: NATO war in 2011 destroyed Libya completely! There should be a break in 2011 in the data, but there is none. To see this more clearly, it is useful to focus on the later period in a separate graph:



WHY is there no break in the data? Why does the infant mortality keep going down, even though all infrastructure was destroyed, all the social welfare systems of the country broke down, and civil war rages through the country for several years? To answer this, we must look more carefully at where the data comes from and how it is computed. Studying the notes in the WDI, we find that they rely on models to compute IM when real data is not available. The nature of these models is not explained. Simple Models which extrapolate on the basis of past trends would lead to such FALSE results.

The fifth country is Bahrain which shows rapid improvements from 1960 to 1975, arriving from a very poor rank to the first quartile in 1975. Thereafter, it follows the first quartile curve. This means that rapid changes in rank occurred in the early years, after which Bahrain followed general trends driving global decline in IM. Only the period 1960 to 1975 would be worthy of study to isolate the key health measures taken by Bahrain to achieve such rapid progress.



What are the lessons that we learn from this data analysis?

1. Data provide us with CLUES about REAL WORLD. The shape of the curves, and the rankings, inform us about the unknown real-world variable – effectiveness of health services – which is qualitative and not directly measurable.
2. Knowledge comes from FOLLOWING up these data clues by the study of complex, multi-faceted, and qualitative real-world phenomena.
3. Just as the data provides clues to the real world, so the real world provides clues about data. Knowledge about destruction of Libya in the NATO invasion in 2011 and the war-torn Yemen suggest that IM statistics for these countries should worsen in the recent years. Since the data show a continuously decreasing trend, we must question the data sources, and ask how these numbers are computed.

4. DATA, by themselves, are NEVER enough to provide knowledge. Knowledge is not knowledge of the numbers, but knowledge of the real world. This is exactly the opposite of conventional statistical methodology, which tells us to restrict analysis to the numbers, and not go beyond the numbers to the real world.

Concluding Remarks

We conclude this lecture with some general remarks comparing conventional Fisherian statistics with the Real Statistics approach, based on Islamic ideals of useful knowledge. We note a few contrasts.

- The main goal of Fisherian Statistics is to REDUCE the data to a few numbers. Once we have these SUFFICIENT Statistics, we no longer NEED the data.
- In contrast, in Real Statistics: EVERY data point is INDIVIDUALLY significant and important.

This lecture illustrates how doing rankings requires consideration of the purpose of the ranking. Different criteria are suitable for different purposes. We showed that use of absolute difference in IM or rank difference helps in finding high performers to emulate for the poorer countries. In contrast, the proportional criterion may be helpful in identifying useful strategies for the high performing wealthy countries.

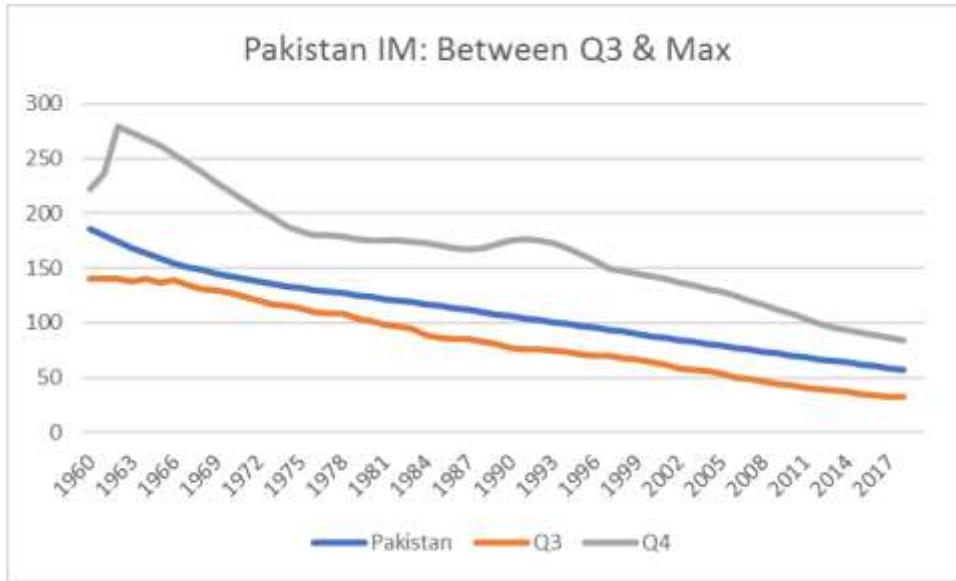
This lecture also illustrates how the data must be considered in conjunction with real world information. Conflicts between the two sources of information may lead us to modify one, or the other, or both.

6F Comparisons, Benchmarks, and Confounding Factors

In the previous part, we compared and evaluated the performance of different countries with respect to reduction in IM (Infant Mortality). In this last part of Lec 6, we focus on the question of which groups are COMPARABLE, when we evaluate relative performance?

First, note that all comparisons are with reference to a benchmark. The benchmarks are subjectively chosen. Benchmarks should be “similar” the country being evaluated, in order to avoid comparing apples and oranges. But SIMILARITY is an elusive concept, and can be used to construct false and misleading benchmarks and comparison. GOAL of this lesson: Provide an introduction to “SIMILARITY”. For similarity, we need to match some factors. The question is: “which factors are VALID to keep matched?”, and “which are NOT valid to match?” for making comparisons. This is best understood in the context of an example.

Using methods of previous lecture (6E), we can evaluate Pakistan’s performance using the following graph:



Here Pakistan's IM is plotted between Q3 and Q4. The graph shows that Pakistan started out above Q3 and stayed above it. There was no change in the ranking of Pakistan, among the 193 countries in the WDI data set. From this we conclude that Pakistan has followed global trends of reduction in IM, neither doing better, nor doing worse. These benchmarks seem fair, since they include all countries in WDI data base. However, one can find & justify alternative benchmarks. For example, a rosier picture of Pakistan's performance would be obtained by removing the wealthy countries from the pool. Which of the two pictures is more accurate? Is it fair to compare Pakistan with much wealthier countries? We will try to answer these questions, later in this lecture.

Before we proceed, we note a common fallacy. One could promote the performance of Pakistan by saying that it has made great progress in IM, going from IM=190 in 1960 to IM=60 in 2018, reduction of over 300%, and saving 130 lives per 1000. This type of statement suffers from lack of benchmark. There is no one we are comparing to. If we say that a student performed well, scoring 80%, this is true only if other students did worse. If all other students scored ABOVE 80%, this would be a bad score. Statisticians should learn to LOOK carefully at the benchmark, also to be able to spot a MISSING benchmark.

Coming back to our main theme, we consider the issue of how to decide “which countries are SIMILAR to Pakistan?”. Depending on how we define similar, we can make the performance of Pakistan look good or bad. Intuitively, it makes sense to remove RICH countries from the picture – they have greater wealth available for the reduction of IM, and they start at a better place. But it does not make sense to remove other POOR countries, like Cuba, on grounds that it is in Latin American, and hence not comparable. How can we sharpen this intuition? What are the FACTORS on which we should have SIMILARITY, to make other countries “comparable” to Pakistan? This is very important and very confusing question, which has only been solved recently, due to deep study of causality, initiated by Judea Pearl in the past few decades. Below we will provide a simplified and intuitively plausible version of the answer generated by this research on causality.

There are TWO conditions on factors for ‘similarity’ to Pakistan:

1. Factor should be EXOGENOUS – it should not be under control of Pakistan, and should not be influenced BY changes in the Infant Mortality rate.
2. Factor should INFLUENCE capabilities of Pakistan to make progress towards lowering IM rate.

These can be summarized in the following RULE: *If factor F affects the capabilities of Pakistan in terms of lowering infant mortality, and if Pakistan can do nothing to change factor F, then comparisons should be made with countries similar to Pakistan with respect to factor F.* This is best understood by looking at some specific examples.

EXOGENOUS FACTORS: We first consider some Exogenous factors. Suppose we say that we should only look at countries in ASIA, as these would be similar to Pakistan, for a fair comparison. This meets the criterion of Exogeneity – Pakistan cannot choose to be in or out of Asia. However, it fails the second criterion. Being in Asia has no effect on the capabilities of Pakistan to lower Infant Mortality. Thus, this is NOT a valid factor for similarity, and confining attention to comparisons with Asian countries would lead to a biased evaluation.

Next consider the factor of WEALTH. Should we exclude Rich countries, with significantly higher levels of wealth, from comparison with Pakistan? According to our criterion, wealth qualifies as exogenous. Pakistan cannot CHOOSE to become rich, at least in the short run. ALSO, being rich DOES increase the capabilities to lower IM. So this meets both criteria, and is a valid factor for similarity. We should compare Pakistan with countries at similar levels of wealth, *at least in the short run*. We will consider the case of the long run later.

Another factor of interest is the Urban/Rural Divide. In the short run, Pakistan cannot change the proportion of people living in rural areas. This also effects the capabilities of lowering IM, because it is harder to get health care to rural areas. Exogenous in the short run. So it would be valid to compare Pakistan to countries with similar ratio of urban to rural population, or to otherwise adjust for this factor.

ENDOGENOUS FACTORS: Next, we consider some endogenous factors. For example, consider Health Spending. Suppose we say that we should compare Pakistan with other countries which have similar levels of Health Spending. That would unfairly restrict competition to countries which do not choose to spend much on Healthcare. Since Pakistan is free to choose how much it spends on Healthcare, this would not be a valid factor for similarity.

Similarly, consider the MECHANISMS by which Pakistan can lower IM. These would include

1. Building more hospitals
2. Training more doctors
3. Public Campaigns to change social norms, to encourage more pre-natal care, and medical attention to pregnancy.

It would not be valid to consider these factors for similarity. Countries spending more on these mechanisms are likely to achieve superior results. Since Pakistan can and should spend more on these, in order to improve its performance, conditioning on these factors will only serve to hide poor performance.

We now return to the question of the exogenous factors Rich versus Poor and the Urban/Rural proportion considered earlier. It is true that these are exogenous in the short run – Pakistan cannot change them. But In the long run, over a period of 50 years, many countries in the sample went from being behind Pakistan to ahead of Pakistan in terms of GNP per capita. This means that it is possible for Pakistan to change this condition in the long run, and so it would be unfair to exclude rich countries, when the period of evaluation is more than 50 years. Similarly, for exogenous factors like urban/rural divide, 50 years provides sufficient time to devise programs to overcome obstacles created by such difficulties. So evaluation against all countries does seem like the correct procedure over the long run.

Confounding Factors in Treatment/Control

The concepts under discussion have much broader application than the evaluation of Pakistan's performance with respect to IM. An important context in which this problem arises is when we wish to assess the value of a medical procedure. As a specific example, consider patients have heart disease. Suppose they are divided into two groups. One group is treated by surgery, while the other group is not. The standard terminology in this situation is that the "Treatment Group" is the one which receives the treatment (surgery), while the Control Group is does not (no surgery). We measure the outcome as 3-year survival rate. Now suppose Treatment group has 70% survival, while the Control group has 50%. Does this mean that surgery is successful? On the surface, it appears to be so, since the treatment leads to greater survival. But appearances can be deceiving.

The CRUCIAL QUESTION is: Are treatment and control groups comparable? The answer goes along lines we have already discussed earlier. F is a confounding factor if it is [1] exogenous – NOT affected by treatment – [2] it AFFECTS probability of survival. If treatment and control are SIMILAR for all confounding factors, then the two groups are comparable, and we can rely on the survival rates to conclude that the surgery helps prolong lives. However, in actual practice in such situations, there is a common problem that arises, which we discuss further below.

Suppose that 100 patients have heart condition. Doctor chooses 50 eligible for surgery – these are TREATMENT group. The remaining 50, not chosen for surgery, are the control group. Suppose we find substantially higher 3-year survival in Treatment group. There can be a reason for this, which creates a difference between the treatment and the control group. It turns out that surgeons choose patients to be eligible for surgery when they don't have additional complicating factors. This means that patients judged as eligible for surgery are healthier than the control group, which generally suffers from additional health problems, which makes them ineligible. In this situation, we say that general health of the patient is a confounding factor. Note that "general health" satisfies the two conditions. It is exogenous – the general health of the patient is a

condition prior to the surgery, which is not directly affected by the surgery. Furthermore, general health does affect the outcome of the surgery. In fact, surgeons select patients in better health precisely because they would like good outcomes for the surgery.

How can we tell if the control group and the treatment group are matched on the confounding factor of “general health”? The solution is to look at all of the patients judged as eligible for surgery, and treat only half of them. AMONG the 50 patients chosen for surgery, randomly choose 25 for treatment by surgery, and leave the remaining 25 as CONTROLS. This time, the two groups are matched on the factor of being chosen by the doctors as being eligible for surgery. When this procedure was actually done in an example described in Freedman’s Statistics (see ‘the Portacaval Shunt’), the results were surprising. Both the 25 patients in the control group and the 25 patients in the treatment group had nearly equal 3 year survival rates, showing the surgery did not prolong lives. However, the original control group of 50 patients considered not eligible for surgery had a much lower survival rate. This shows that the poor general health of the patients considered ineligible for surgery WAS the cause of their lower survival, and NOT the failure to get surgery. The wrong conclusion obtained initially was because the treatment and control group were not matched with respect to the confounding factor of “general health”.

Seeing several illustrations of confounding factors helps to understand the issue better. So we consider a second illustration of an Expensive Drug for some disease. It is noted that among patients who take drug, recovery rate is 80%. Among OTHERS, who do not take the drug, the recovery rate is 50%. Can we conclude that the drug is effective? Are treatment and control comparable? It is possible that those who can afford the drug are WEALTHIER and have better health. Those who cannot afford to take the drug are POORER and have worse health. In this case, the lower recovery rate among those who do not take the drug is due to their worse health status, and not necessarily due to the drug. The solution is to run an experiment which compares the rich with the rich and the poor with the poor. Note that WEALTH is EXOGENOUS, and affects outcomes.

A third example comes from initial trials of Salk’s Polio Vaccine in the USA. The vaccine offered to school children on a voluntary basis. Those who ACCEPT are treatment group, and those who REJECT are control group. Studies of the treatment and control group revealed that those who accepted were, on the average, wealthier than those who rejected. The poor had a greater tendency to distrust unproven vaccines than the wealthy. Now Polio is unusual in that the wealthy are MORE vulnerable to it than the poor. Because of poor hygiene, the poor are more likely to be exposed to the virus, and hence develop immunity, relative to the rich. Because of this difference in the treatment and control groups, wealth is a confounding factor. Initial experiments like this had outcomes where the Poor in the control group showed relatively good results without vaccine because of their greater immunity. The Rich in the treatment group showed good results with vaccine. But the effect of the vaccine appeared to be reduced because of the confounding factor. Realizing this source of bias, statistician devised better experiments to match the treatment and control group with respect to the confounding factor. See Freedman for details.

Our final example comes from a Prison Recidivism Study, also discussed in Freedman. The USA has the largest percentage of prisoners (as a proportion of its population) on the planet. When prisoners are released from jail, they often end up coming back to jail for some other crime some time later. This is called “recidivism”. In an effort to reduce this return to prison, an intensive and difficult program was devised for training prisoners, in the hope that this would give them the skills to stay out of jail. Prisoners who were interested in the program could volunteer to participate, but no one was forced to take the program. After an year or two, the recidivism rate was compared between the volunteers who took the training, and the rest who did not. Sure enough, the volunteers had lower rates of return to prison. Does that prove that the training program was successful in lowering rates of recidivism?

We must check for confounding factors which create differences between the control and treatment group, and also affect the outcomes. One such Confounding Factor is MOTIVATION. Those who volunteered for the difficult program were obviously more motivated to work hard to stay out. Those who did not volunteer had less motivation. It is clear that this factor is Exogenous (because it caused prisoners to volunteer) and it would obviously affect the rates of recidivism. Therefore it is a confounding factor. We cannot tell if the lower rate of recidivism is due to the training program or just due to the additional motivation of the volunteers.

There are two ways to overcome this bias. If the treatment and control groups are randomly chosen from among the entire population of prisoners, then the two would be comparable. Some unmotivated prisoners would be forced into taking the training, while some motivated prisoners would not be allowed to take the program. In this case the two groups would be comparable, and we could come to the right conclusions. If we do not want to force unwilling prisoners to take the training then there is an alternative SOLUTION. We can divide the VOLUNTEERS into two groups – Treatment and Control. NOW both groups would have equal motivation, since both groups volunteered. Comparing these two groups with respect to recidivism would give us the right results regarding the effects of training. This is because the treatment and control group are matched with respect to the confounding factor of motivation.

Concluding Remarks

We review some of the key points made in this lecture. The first is that whenever we evaluate performance, we have an explicit or implicit benchmark in mind. Those who being evaluated (the treatment group) must be compared to the benchmark group (the control group) to arrive at a conclusion regarding their performance. For a fair comparison, it is essential that the two groups should be MATCHED on the confounding factors. Confounding factors are those which are exogenous and affect outcomes. The groups SHOULD not be matched on endogenous variables. Also, they should not be matched on exogenous variables which do not affect the outcomes.

Here, the definitions of Exogenous and Endogenous, and how we can tell if the factor affects Outcomes depends STRONGLY on our knowledge about the real world. The theory is more complex than the simplified and intuitive picture we have described here. However, the main point of importance here is that we cannot learn about the effectiveness of a treatment

without understanding and adjusting for these real world effects caused by confounding variables. Statistician have remained confused about this issue for decades because they have searched for solutions within the data, without reference to the real world. In particular, Big Data CANNOT resolve these issues, since the causal patterns required to define exogeneity and endogeneity are not part of the data at all. Instead, they form a part of the (unobservable) causal mechanisms which operate in the real world. Causal mechanisms are unobservable because when we say the X caused Y, we also imply that if X had not happened, then Y would not have happened. This is something we can never observe. This point was noted by philosopher David Hume a long time ago when he remarked that we can only observe that Y occurred after X; we cannot observe that X caused Y.

7: Probabilities, Binomials, and p-values

The Chapter Blurb – to be added

7A: A New Definition of Probability

Lecture 8A provides answers to the question of “What is Probability?”. An article by [Alan Hajek in Stanford Encyclopedia of Philosophy](#) lists six major categories of definitions. Many more are possible if causality is also taken into account. These definitions conflict with each other, and face serious problems as interpretations of real-world probabilities. The basic definition of probability we will offer in this lecture falls outside all of these listed categories. Before going on to present it, we briefly explain why there is such massive confusion about how to define probability.

1 Emergence of Probability in Europe

According to Ian Hacking’s account in the Emergence of Probability, modern concepts of probability emerge in the middle of the 17th Century Europe, and have not evolved substantially since then. He explains that these formulations were shaped by (now forgotten) historical contexts which constrained the space of possible theories. This explains why satisfactory definitions of probability are not available even in the beginning of the 21st Century. To vastly oversimplify, we provide TWO major reasons why attempts to define probability went astray.

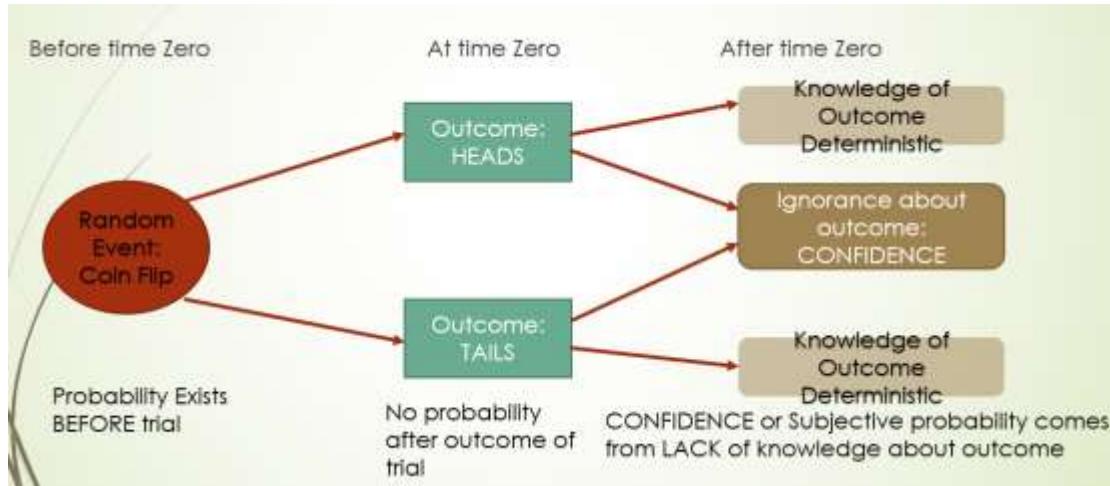
ONE: Colin Turbayne, in his superb but neglected book [“Myth of Metaphor”](#), has documented how Newtonian metaphor of the world as a machine, governed by deterministic laws, replaced an earlier metaphor of the world as an organism. In a deterministic world, there are no chance events. This constrains probability to be an aspect of our ignorance regarding the laws governing the universe. This precludes the existence of probabilities in external reality.

TWO: Empiricist philosophies originating with Hume and culminating in Logical Positivism deny the relevance of unobservables to science. Probability is intrinsically concerned with what might have happened, and hence not definable in terms of what did happen.

Volumes can and have been written expanding on these brief remarks. For my own detailed exposition, see: [Subjective Probability Does Not Exist](#).

2 A New Concept of Probability:

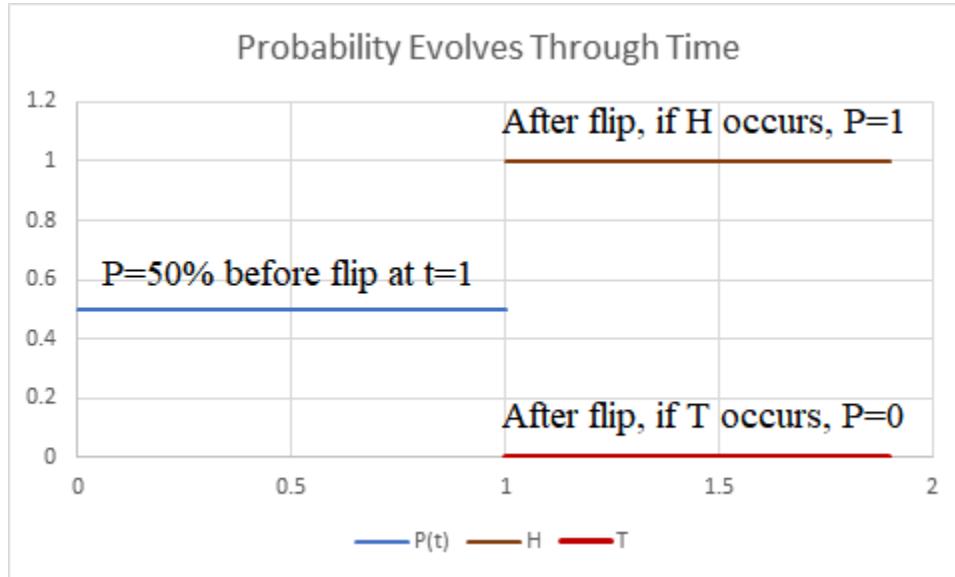
We will now define probability via a new framework, for use in this course. Consider a random event such as a coin flip. Our definition is Chronological – it takes into account a central feature of probability not available in the major definitions. Probability exists and is well-defined BEFORE the coin is flipped.



We understand this probability as a “propensity”, a feature of external reality having to do with the coin and how it is flipped. After the flip, an outcome occurs – normally either Heads or Tails. When the outcome occurs, the probability is EXTINGUISHED. We are now in one of two possible mutually incompatible worlds, as depicted in the diagram above. In either of the two worlds, we cannot talk about the Probability of Heads or Tails, because the uncertainty has been resolved, and a definite outcome has occurred. However, suppose that I do not know which of the two outcomes occurred. In this case, my knowledge of the pre-event probability gives me information, which can be formalized using the term “CONFIDENCE”. We can say that I have confidence level of 50% that Heads occurred, and similarly for Tails. This confidence is subjective – personal to me – because of my ignorance about what happened. Someone who knows what happened would not share my personal assessment at all. Confidence comes into being after probabilities are extinguished, and does not exist prior to the occurrence of the random event. To be more precise, there can be no significant differences between subjective and objective assessments of probability prior to the occurrence of the event. All such evaluations would be equally subjective and/or objective.

To express this formally, use X to denote a random variable, representing a coin flip. Traditional notation writes $P(X=\text{Heads})=P(X=\text{Tails})=1/2$ for fair coin. However, this ignores the temporal nature of probability. For each random variable, there is a time of OCCURRENCE, a time at which the random variable takes a particular outcome and uncertainty is resolved. This must be an essential part of the description of the random variable. Let us write $X[s]$ for a random variable which is uncertain upto time s . Then the coin flip occurs at time s and results in

a specific outcome: Heads or Tails. In talking about Probability, we must also mention the time at which probability is being evaluated, since this probability changes with time. Let us use $P(t)$ to denote probability at time t . Then the simplest picture of the probability of a coin flip is depicted in the following function: $P(t)\{X[s]=\text{Heads}\}$:



Here, the coin is flipped at time $s=1$. Prior to $t=1$, probability of Heads is 0.5. After $t=1$, this probability is either 100% when Heads occurs as an outcome of the coin flip, or 0%, when Tails occurs.

This definition incorporates several features not available in any of the current major definitions of probability. As explained by Hacking, conceptions of probability were constrained by historical context, and have not succeeded in liberating themselves from these original constraints. In particular, understanding of causality is closely connected with alternative possible worlds, which are different from, and yet similar to, our experienced reality. In this conception of probability, the coin flip creates two possible worlds, which differ from each other only in this single respect – in one world, the coin came out heads, while in the other it came out tails. As time progresses, differences created by this branch may be extinguished, or may be amplified. For example, if an important decision (like choosing strategies or teams) is linked to the coin flip then difference in outcomes could lead to further differences. But all these topics, and connections to counterfactuals and causality, are not relevant to the present lecture. Here we aim to provide an elementary introduction to probability for beginning students.

3 Conjectural Probability

In the picture of probability above, a crucial element is missing. How do we assign probabilities to the two branches? How do we know that Head and Tails are equally likely and have 50% probability? If the event is Rainfall or Clear Weather tomorrow, what are the probabilities to be assigned to the two branches?

As per critical realist philosophy, scientific theories are conjectures about the hidden structures of reality. These can never be verified directly, because these structures remain unobservable. But sufficiently good matches between predictions of these theories and observed outcomes give us indirect confirmation, and a reason to trust these theories. Along these lines, the probabilities ascribed to the branches represent a conjecture about reality, which can never be verified to be true.

According to positivist dogma, a sentence with unverifiable truth value is meaningless. This, and other intellectual predilections among founders of modern concepts of probability made it impossible to arrive at definitions based on alternative unobservable realities.

We now illustrate conjectural probability with examples. To be more precise, conjectural probability is a hypothesis about unobservable structures of reality which generate the observations that we see. This hypothesis can never be directly verified because these hidden structures (which govern what might have happened) can never be directly observed. However, indirect confirmation is possible and occurs when predictions of our theories match observed outcomes.

1. “Fair” Coin

A Coin Toss is often used for choosing at random among two options. The hypothesis of “fairness” is a conjecture that the two possible outcomes are equally likely. The physical symmetric structure of the coin, the circumstances of tossing, and historical experience, all provide evidence for the conjecture that $P(\text{Heads})=P(\text{Tails})=50\%$, prior to the toss. After the toss one of the two outcomes occurs, and probability is extinguished. If a person is IGNORANT of outcome, he can have SUBJECTIVE probability. He can have CONFIDENCE 50% that Heads occurred, and similarly CONFIDENCE 50% that Tails occurred.

2. Equal Skill in a Game

Suppose we CONJECTURE that two players A & B have EQUAL skill. One meaning of “equal skill” is that both have equal chances of winning a particular game. Thus the conjecture of equal skill leads us to assign equal probabilities: $P(A \text{ wins})=P(B \text{ wins})=50\%$ if DRAW is not a possible outcome. If DRAW is possible, then $P(A \text{ wins})=P(B \text{ wins})=50\% \text{ of } (1-P(\text{DRAW}))$. Conjectural Probability created by adding a THEORY about real world and attempting to see if it matches the observations. Note that we may believe the theory to be obviously false. The point of putting down the theory is to reject it by using data, providing an empirical proof that one of the two players is more skillful.

3. Drug Versus Placebo

Take two groups of patients – Treatment & Control Group. Match the patients in pairs, assigning one control to one treatment: T1 to C1, T2 to C2, . . . We attempt to match the pairs on all relevant factors. Give Drug to treatment group and Placebo to control group. We CONJECTURE that the Drug has no effect (Drug equivalent to Placebo). As a consequence, difference in Outcomes is purely random – that is, the two outcomes (T1 cured, C1 not) and (T1 not, C1 cured) are equally likely. We can use predictions of this conjectured probability to assess

whether or not the conjecture is in conformity with empirical evidence. Again, we will normally be interesting in a statistical REJECTION of this conjecture. That would give us empirical evidence for the efficacy of the drug.

4. Drawing Straws:

There are twelve people on a boat which gets caught in a storm. They believe that the storm has been sent to punish a guilty person who is fleeing the scene using the boat. They decide to draw straws to determine who is guilty. Twelve straws of equal length are taken, and one of them is broken off at one end. The break is concealed, and the bundle held out to all parties sequentially. The one who ends up drawing the short straw is deemed to be guilty and thrown off the boat. The probability conjecture is that all straws have equal probability of being drawn. Under this hypothesis, it can be shown that all parties have equal chances 1/12 of being found guilty.

5. Independence and ESP

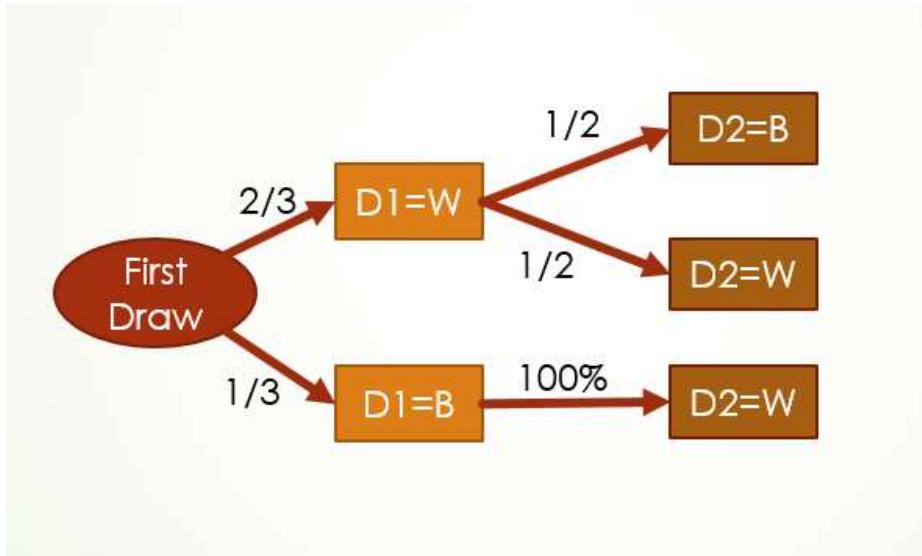
Two people are given a deck of cards. One of them is asked to choose one of them at random, put it aside, and pick a second one, and continue through the deck in this way until the end. The second person is asked to guess the color of the card and record his guess (Red or Black). At the end of the experiment, we will have 52 pairs of the type (R,R) (R,B) (B,R) and (B,B). The conjecture that the guess and the draw are independent leads to equal probability for all 4 pairs. In contrast, if the second party has ESP, then (R,R) and (B,B) will occur more often than they should under independence. We can check the data to see if it is in strong conflict with the conjecture of independence, as evidence for ESP.

6. The Vietnam Draft Lottery

The Vietnam Draft Lottery was instituted to create a fair way to choose how to send young men to kill Vietnamese, and possibly die in the process. This was done by randomly choosing from among the 366 days of the year, and also randomly choosing between the letters of the alphabet A,B,...,Z, to decide on ranking of young men, according to birthdate and initial letters of the names. Under the conjecture that all dates and names had equal probability, all of the young men in the target age range would have equal chances of being picked. There was substantial controversy regarding this, and ultimately it was concluded that the process by which the dates had been picked did not give equal chances to all days of the year.

4 The Bayes Rule

Our conception of probability and confidence allows us to resolve a dilemma which has been a source of controversy for centuries. Consider an urn containing three balls, two of which are White while one is Black. At time $s=1$, we make the first random draw D_1 from the box. Thus $P(D_1=W)=2/3$ and $P(D_1=B)=1/3$. Here we adopt the simplifying convention that when time is not mentioned for probability or for the random variables then both times are assumed to be PRIOR to the occurrence time.



At time $s=2$, we make the second random draw. The probabilities for this draw are as shown in the diagram. If $D1=W$ then remaining balls are WB and both have equal chances of being drawn. If $D1=B$ then both remaining balls are white, so the only possibility is $D2=W$.

The branches on the diagram leading to outcomes of the second draw $D2$ have conditional probabilities. These are probabilities based on OUTCOMES of $D1$. The uncertainty associated with the first draw has been extinguished, and a particular fixed outcome – $D1=W$ or $D1=B$ – has occurred. Suppose we want to compute the probability of a sequence of draws, such as $D1=W$ and $D2=W$. This probability changes with time. At $t=2$, all probabilities are extinguished: $P(2)\{D1=W \& D2=W\} = 100\%$ or 0% , depending on the outcomes of the two draws. At $t=1$, uncertainty only attaches to the second draw. The outcome of $D1$ has already occurred. There are two possibilities – either $P(1)\{D2=W\}=50\%$, when $D1=W$ occurred, OR $P(1)\{D2=W\}=100\%$ when $D1=B$ occurred.

To understand the Bayes controversy consider calculating the probability at time $t=0$ of a sequence of draws such as $D1=W$ and $D2=W$. It is easily seen that

$$P(0)\{D1=W \& D2=W\} = P(0)\{D1=W\} \times P(1)\{D2=W|D1=W\} = 2/3 \times 1/2 = 1/3$$

Here the conditional probability written as $P(1)\{D2=W|D1=W\}$ has a natural interpretation. At time $t=1$, the first draw has OCCURRED, and an outcome has been observed. All probabilities going forward depend on which outcome occurred, and so this must be specified as the condition, to allow computation of the probability.

The Bayes formula arises when we fail to keep track of time while evaluating probabilities. The formula above is written in traditional notation as $P(A \& B) = P(A) \times P(B|A)$, and this is taken to be a universal law of probability. When B follows A in time, this makes sense, and corresponds to our formula. However, now consider reversing the roles of A and B. This gives $P(A \& B) = P(B) \times P(A|B)$. If A and B are contemporaneous events, this would work fine. But in the present case the second formula can be written as follows, without indicating time:

$$P(D1=W \& D2=W) = P(D2=W) \times P(D1=W|D2=W)$$

The Bayes rule allows the calculation of this probability, according to mechanical rules. Controversy arises regarding the meaning of this calculation. The formulae has the un-natural conditional probability which asks about the occurrence of $D1=W$ given that $D2=W$. At what time is this probability being evaluated? The natural time is $t=2$, when the outcome of the second draw has occurred and is W . The key point here is that knowing $D2=W$ gives us the additional information that $D1$ must have occurred. The second draw cannot be made without making the first draw. Thus, probabilities associated with $D1$ have been extinguished. The chronologically sequenced events are no longer symmetric, and the conditional probability which makes perfect sense in the natural direction of time flow makes no sense when we reverse the direction of time. In particular, we must ask about the specific circumstances in which two draws were made in such a way that the first one was concealed, but the second one was revealed to the observer. In this situation, no observer would make bets with others about $D1$. $D1$ has occurred and while the observer is ignorant about this outcome, others may well have observed it.

5 CONCLUDING REMARKS

Ian Hacking in “The Emergence of Probability” writes that:

The preconditions for the emergence of probability determined the space of possible theories about probability. That means that they determined, in part, the space of possible interpretations of quantum mechanics, of statistical inference, and of inductive logic.

That is, historical circumstances in Europe determined the ways of thinking about probability which emerged – and EXCLUDED many other possible ways of framing thought about probability. Other histories could lead to other concepts of probability, and Hacking mentions tantalizing alternatives from Indian history. In this lecture, we have provided an alternative framework to conceptualize probability, and to resolve the vexing conflict between subjective and objective probability.

7B Rules of Probability

General Introduction: Meanings and Philosophy

In lecture 8A, we gave a new definition of probability. Each random event has many possible outcomes. One of these possibilities is realized, at which point all other possibilities become “what might have happened”, while the realized outcome acquires 100% probability. Probability is about FUTURE POSSIBILITIES. Everything which can happen creates possible future worlds.

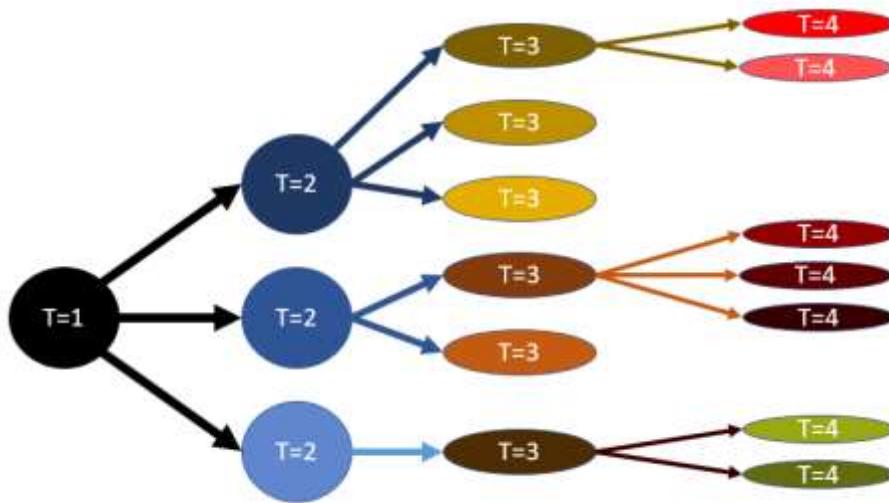
In the early 20th century, there was a huge debate between two different conceptions of uncertainty. On the one hand, Keynes and Knight held that the future was completely uncertain. We did not know the range of possibilities, and we did not know the probabilities to be assigned to these possibilities. As opposed to this, Ramsey and De-Finetti argued that rational decision making requires knowledge of all possible future outcomes, as well as their probabilities. This second conception, that we have knowledge of future outcomes and their probabilities,

eventually won out. Modern theories of decision making under uncertainty rely on the Ramsey De-Finetti approach, while the Keynes-Knight approach has been marginalized. The Global Financial Crisis, and many other events, prove conclusively that theory of rational expectations is wrong. Although of tremendous importance, not topic of our study here.

Probability is a mental MODEL of structures of external reality. Following the Keynes-Knight approach, the future is fundamentally uncertain. We can never know hidden structures of reality, and we can never know the future. We ONLY have models of future possible outcomes & probabilities. We NEVER have “Knowledge” in the sense of JTB: Justified True Belief. The Western intellectual tradition put the bar for knowledge too high to allow for the definition of probability that we are using here. Our models are just best guesses based on our experience, and will forever remain unverifiable. There is subtle and fine distinction between subjective & objective models here. Even though it is important, we will not discuss these philosophical aspects pertaining to the meaning of probability further. Rather, we will work with probability in cases where there is substantial consensus about its meaning, and avoid controversy.

Probability Models are Trees in Time

We define a probability model to be a tree which grows branches as time progresses:



A probability tree, like the one above, consists of nodes and branches. Each branch has a probability. Each node is a possible outcome, situated in time. Probabilities of nodes vary with time. At any node, the sum of probabilities going forward from that node should equal 100%. That is, the set of nodes at the next time level should cover all possibilities – and one of these possibilities must happen. If we go forward two steps in time, then the probability of a node is obtained by multiplying the probabilities of the two branches which lead to the node.

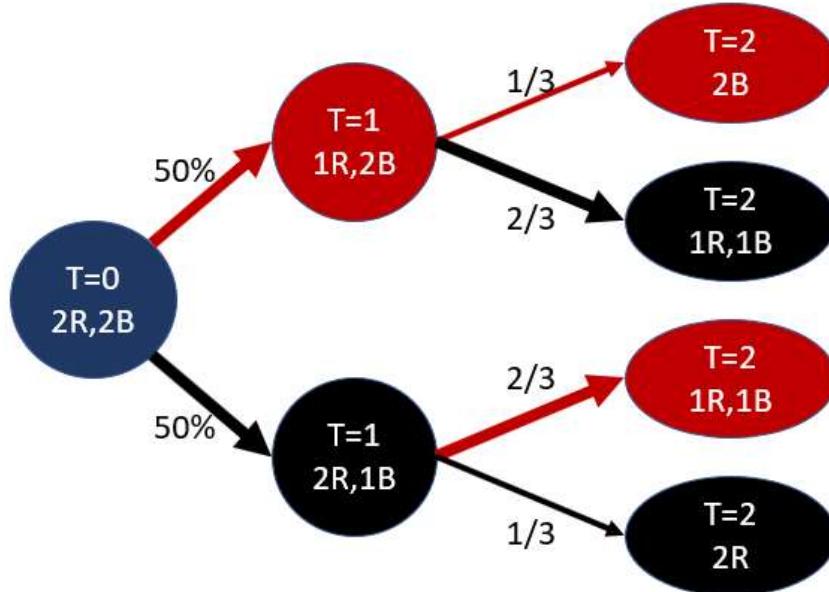
The most important feature of this model is systematically time-varying probabilities. In the above diagram, at T=1, the three blue circles are in the future, and have probability weights given on their branches. At T=2, one of the three events OCCURS. At this point, the other two events become unrealized alternatives, worlds which could have been, but never will be. Beyond this time, the event which occurs has 100% probability. All alternatives on all other branches have become impossible, and their probabilities are now set to 0%.

Outcomes: An Outcome is a Node, which is positioned at a specific time T, on the time-branching probability tree. From time T onwards, the outcome which occurs at time T has probability 100%, and all others have probability 0%. At time T-1, the probability of the outcome is the probability assigned to the branch which leads to this outcome. At time T-2, the probability of the outcome is the product of the two branches which leads to the outcome. And so on.

Events: Events are things which can happen in multiple ways. Thus there are many different nodes at which the even occurs. Probability of future event is SUM of probabilities of all possible ways the event can happen. That is, we compute probabilities of each of the nodes at which the event occurs, and add these probabilities to get the probability of the event. Nodes on different branches of the tree are mutually exclusive – if one happens then the other one cannot. Probabilities can be added for such nodes. If two nodes are along the same branch, then one cannot add their probabilities. Rather, one must determine the first node at which an event occurs on a given branch, and add this probability to that of the other nodes.

We now illustrate the use of these rules of probability in some simple examples.

Drawing Balls from Urns: An urn contain 4 balls. Two are red, and Two are black. We make two draws at random from these urns. The time-branching probability graph below is a probability model for this situation:



At the first draw, there are equal numbers of red and black balls, so there are two possible events, black or red, and both have equal probability. At T=1, if R was drawn, probability of the 2nd draw of R is now 1/3, while B draw is 2/3. This is because after a black draw, there is one red and two black balls in the urn. Similarly, After initial black draw at T=1, $P(T=2:R)=2/3$ and $P(T=2:B)=1/3$. It is essential to use the time index, since probabilities change with time, depending on which outcome occurred in the past.

Now consider asking the probability at time T=0 of a Red ball at time 2: $P[T=0]\{T=2:R\}$. There are two outcomes with a red ball at T=2. We can write them as $[T=1:R] \Rightarrow [T=2:R]$ and $[T=1:B] \Rightarrow [T=2:R]$. Here the \Rightarrow indicates the time sequencing and can be read as “followed by”. Draw of R at T=1 followed by another R at T=2 has probability equal to $1/2 \times 1/3 = 1/6$, by multiplying the branch probabilities. We can write this symbolically as $P\{[T=1:R] \Rightarrow [T=2:R]\} = 1/6$. Similarly the probability of Black at T=1, followed by Red at T=2 can be computed by multiplying the branch probabilities: $P\{[T=1:B] \Rightarrow [T=2:R]\} = 1/2 \times 2/3 = 2/6$. There are two ways to draw Red at T=2, and we ADD these two probabilities to get the desired probability:

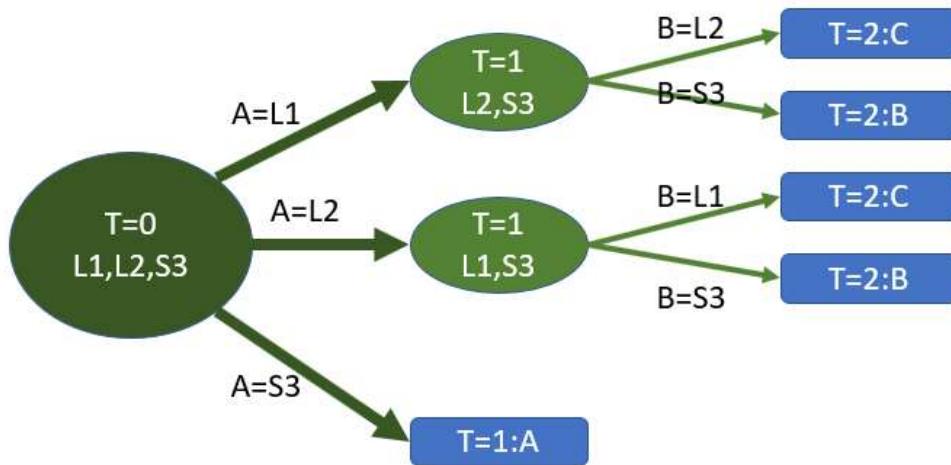
$$P[T=0]\{T=2:R\} =$$

$$P[T=0]\{[T=1:R] \Rightarrow [T=2:R]\} + P[T=0]\{[T=1:B] \Rightarrow [T=2:R]\} =$$

$$(1/2 \times 1/3) + (1/2 \times 2/3) = 1/6 + 2/6 = 1/2$$

Drawing Straws: As a second example of how we create probability models as branching trees, and use them to compute probabilities, we consider the “drawing straws” example discussed in the previous lecture on the definition of probability. There are N people and N straws. One straw is short. All others are long. Straws are held so that the short end is

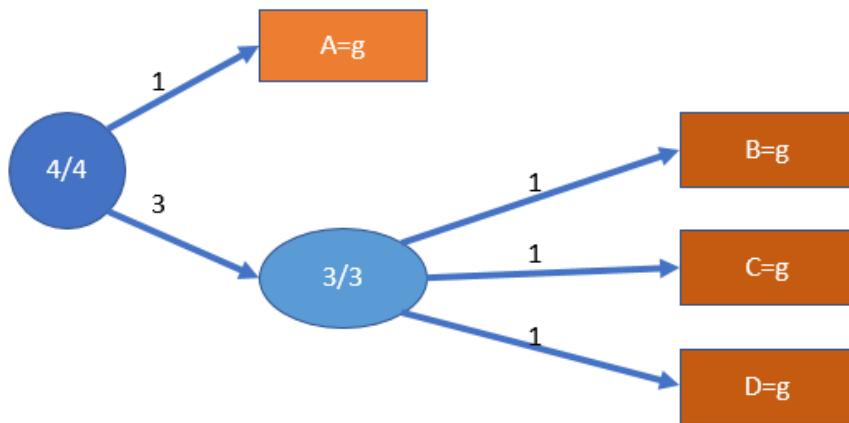
concealed. All straws look alike to the person who is drawing. People draw straws in sequence. Are the early people MORE likely to draw the short straw? OR are they LESS likely to draw the short straw? Or, are the probabilities the same for all people? To answer this question, we do the calculations. As a general rule, it is best to start with simple cases. This builds understanding. The case of two people and two Straws is trivial; it is left for the student. We go to the next case of 3 People A,B,C with 3 Straws. Label the three straws as L1, L2, S3; here L1 and L2 are the two long straws and S3 is the third short straw. A probability model for this situation is pictured below:



Probabilities change across time. At T=1, there are three possible branches. On one branch A has been chosen with 100% probability. On the other two branches, A has chosen the long straw and has been eliminated. That is, the probability of A being chosen is now 0%. Also, at T=1, we cannot ask the question of what is the probability that B will draw the short straw and be chosen ($P[T=1](T=2:B=S3)$), WITHOUT specifying what happened at T=1. First draw probabilities have been extinguished. We must know which node we are on, in order to calculate the probabilities going forward into the second stage draw. These probabilities can be called **CONDITIONAL PROBABILITIES** – that is, we must specify the **CONDITION** (what happened at T=1) in order to compute future probabilities. This conception of conditional probability as

chronological, referring to sequencing in time, is new to this definition, and different from classical definitions.

Four or More Straws: Once we have solved the 3/3 case, with 3 persons and 3 straws, it is easy to solve the 4/4 case. Suppose we have 4 people (ABCD) and 4 straws, with one short straw. We already know the probabilities in 3/3 case. So we can draw the probability model for the 4/4 situation as below. At the first draw, at $T=0$, there is 1/4 chance that A is chosen. In the other 3 out of the 4 possible draws, A is eliminated, and also one straw is removed. This means that we are back to the 3/3 case, with 3 persons and 3 straws. This case we have already solved and determined that each of the 3 people have equal chances 1/3 of being chosen. It is similarly easy to go forward and show that the same holds for any number N, with N persons and N straws.



Housing Lottery: When I was an undergraduate at MIT in the early 1970's, there was a wide variety of housing choices, and some were more popular than others. In order to equitably distribute housings, students were randomly assigned a number which would determine their priority in housing choice. The student with ticket #1 would get to choose first, #2 would choose second and so on. To model this situation, suppose that the Housing Office has tickets marked from 1 to 1000, one for each student. Students show up at random. Each one is given a ticket chosen at random from the ones which remain. After all tickets are assigned, students choose housing in sequence according to the ticket number. It used to happen that students would line up in the morning to get the early tickets, because of fear that the good numbers will be gone by the afternoon. Is this fear justified?

Our analysis of the short straw enables us to answer this question. Think of Ticket #1 as the short straw. All people have same probability of drawing short straw. Thus, all students, regardless of when they show up at the housing office, have equal probability of drawing Ticket #1. Similarly for ANY ticket in Housing lottery, all students have equal chances of drawing THAT ticket. Thus all students have equal chances for drawing all tickets.

Concluding Remarks: The central feature of our model is that probability exist ONLY for Future Events. This lecture was mainly about the rules for calculating probabilities when they are modelled by time-branching trees. The simple rules can be summarized as follows. At each

NODE, probabilities on branches going forward are CONDITIONAL on getting to that NODE. Probabilities going out multiple steps forward are calculated by multiplying probabilities on branches. Probabilities for Events which occur on multiple branches can be obtained by ADDING probabilities of all the nodes on which the event occurs. As time advances, some probabilities are extinguished, and branches corresponding to those possibilities get removed from the tree. These are things that might have been, but are now forever impossible. This time varying feature is an essential aspect of probability, but is not captured by current models of probability.

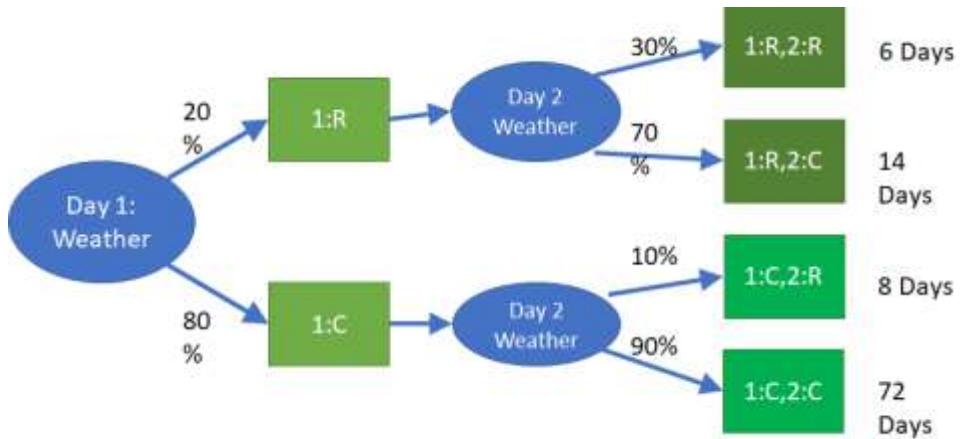
Exercise: To get practice in applying these rules, we describe a simple probability problem. Remember to solve for simplest cases first - N=2 is trivial, N=3 is simple.

An airplane has N seats and N passengers. Each passenger has been assigned one seat: P1 => 1, P2 => 2, and so on, PN => N. The FIRST passenger ignores his assigned seat, and sits down at random on any one of the seats 1,2,...,N – ALL seats have equal probability on this first choice. All other passengers go to their ASSIGNED seats. HOWEVER, if the seat is already occupied, they choose a seat at RANDOM from all unoccupied seats. What is the probability that the LAST passenger gets to sit in his ASSIGNED seat?

7C Applications of Time Branching Probability Models

In previous lectures, we have developed time-branching models of probability. In this section, we show how these models help clarify concepts and resolve outstanding puzzles. We first look at how probability models can be used in the context of weather forecasts.

As a hypothetical example, we construct a probability model for a sequence of two days. We consider only two possible events, R (rainfall) or C (clear). The tree of possible outcomes can be pictured as follows.



We have created a hypothetical model to demonstrate dependence. In particular, rainfall is more likely to follow rainfall, and a clear day is more likely to follow a clear day, according to this model. It is useful to think of the model as being about 100 potential first days. On 20 of these alternative reality days, it will rain, while on 80 of them it will be clear. Going on further,

on the 20 days that it rains, 30% of the time ($30\% \times 20 = 6$) it will rain again, while 70% of the time (14 days) it will be clear.

The mathematical notations to express the above information in Time-Dependent Probability models is awkward, because we need to specify both the time at which probability is being evaluated, and the timing of the events for which we are evaluating probability. We can write that $P[T=0](1:R)=20\%$. This says that the Probability [At time T=0] of the event 1:R (rainfall on the first day) is 20%. To simplify the notation, we can adopt the convention that when time is not mentioned, then T=0. Also T=0 is always the time before any random events take place – all events occur at some positive time.

With this convention, $P(1:R)$ refers to the probability at T=0 that rainfall will occur on the first day, which is 20%. We can also write $P[T=1](1:R)$ for the probability of rainfall on day 1, when day 1 has occurred. In this case, the probability is either 0% (when rainfall does not occur) or 100% (when rainfall occurs) on the first day, and on all subsequent days.

Next, consider evaluating the probability of two days of rainfall. We can write this as $P(1:R=>2:R)$. Since time has been omitted, the probability is to be evaluated at T=0. We use 1:R => 2:R to indicate rainfall on 1st day followed by rainfall on 2nd day. This notation captures the time sequencing of the two events. In this notation, it is worth noting that we cannot evaluate $P[T=1](2:R)$. This is because the probability of rainfall on the second day depends on what happens on the first day. So, $P[T=1](1:R => 2:R)=30\%$; this specifies that 1:R occurred at T=1. Similarly, $P[T=1](1:C=>2:R)=10\%$; if the first day was clear, then probability of rainfall on second day is reduced to 10%. After developing all this notation, we can now write the formula for the probability of rainfall on both days as:

$$P[T=0](1:R => 2:R) = P[T=0](1:R) \times P[T=1](1:R => 2:R) = 20\% \times 30\% = 6\%$$

This formula is much simpler in standard notation, which does not have time-sequencing. For any two events A & B, we have:

$$P(A \& B) = P(A) \times P(B|A) = P(B) \times P(A|B)$$

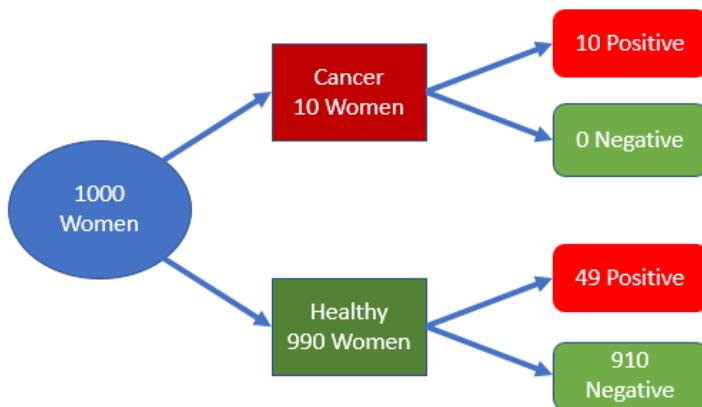
But Time-Reversal is NOT possible in time branching models. The Bayes formula is based on doing the reversal, and does not hold with time-branching probabilities.

There is a very important special case of INDEPENDENCE, where the probability of events on the 2nd day does not depend on what happens on the first day. Our original model has been setup to show dependence. $P[T=1](1:R=>2:R)=30\%$ Evaluated on Day 1, after Rainfall has occurred, $P[T=1](1:C=>2:R)=10\%$ Evaluated on Day 1, after Clear Day. Probability of rainfall on Day 2 DEPENDS on weather on Day 1. INDEPENDENCE holds when both probabilities are the same. $P(2:R)$ is fixed regardless of what happens on Day 1. In this case: $P[T=0](1:R=>2:R) = P[T=0](1:C=>2:R) = P[T=0](2:R)$. In this situation, it becomes possible to write $P[T=1](2:R)$ – we do not need to specify what happened on day 1, because $P(2:R)$ is independent of outcomes on Day 1. In the case of independence, we get the following formula for probability of rainfall on both days: $P[T=0](1:R=>2:R) = P[T=0](1:R) \times P[T=0](2:R)$. Using the convention that T=0 can be omitted, we can simplify this to: $P(1:R=>2:R) = P(1:R) \times P(2:R)$

Given a real -world situation, how do we know which model to apply? Are weather events on subsequent days independent or not? There are two sources of information we can use to build such models, and answer such questions. One source is empirical evidence based on past rainfall data. The other source is based on physical mechanisms which generate weather. That is, configuration os wind, clouds, atmospheric pressure, and other factors integrated into weather forecasting models would allow us to generate probability models for weather. In general, we take both of these types of information into account in constructing probability models.

The Base-Rate “Fallacy” Explained:

In this section, we will explain what the base-rate fallacy is, and why it creates controversy. We will explain it in the context of the canonical example of screening for breast cancer. The BASE RATE is the proportion of females in the population who have breast cancer. Let us suppose, to begin with, that 10% of Females in high-risk age group have Breast Cancer. We have a medical test which screens for cancer. Screening detects 95% of these case. That is, SENSITIVITY = 95%, which means that if a woman has breast cancer, screening will detect it with 95% probability. But, screening will also “detect” cancer in healthy woman. That is, SPECIFICITY=90%, which means that screening leads to false positives in 10% of healthy women. The question we want to ask is: Given a positive outcome (test detects cancer), what is the chance that woman has cancer? Drawing a time-branching diagram gives clarity about the issues involved. As we can see, using the base rate of 10%, 100 women will have BC in a population of 1000. Among these, women, screening will detect 95% or 95 women. But, the screening will also “detect” 10% of the 900 healthy women, falsely tagging 90 of them as being positive for BC. So the probability of cancer among the women who are screened positive is 95/185 or approximately 50%.



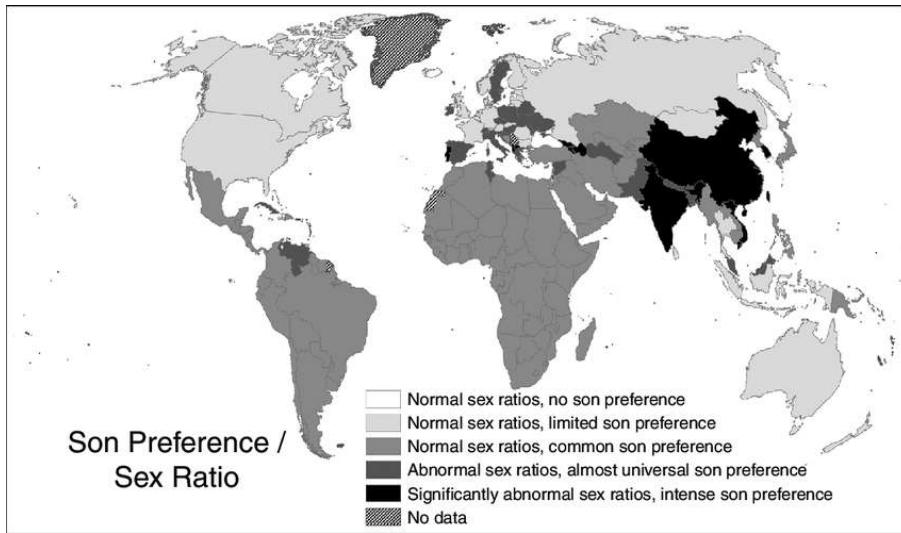
If we change the base rate to 1%, then only 10 women will have cancer, while 990 will be healthy. The test will detect 99 healthy women to be positive. Thus, the probability of cancer given a positive outcome is only 10/109 or around 10%. This shows the strong influence of base rates on the answer to the question of the probability of cancer given a positive screening result. YET, most people disregard the base rate when giving their views about this probability. This is called the “base-rate fallacy”.

DISCUSSION: According to our time-branching probability model, the base rate probabilities are valid BEFORE the screening. After a particular woman is selected and screens positive, there are no probabilities involved in the question: does she have breast cancer? The answer is either YES or NO. Most people intuitively realize that the base probability does not apply once a particular person has been selected. To see this, note that any particular woman is a member of many different populations. She may be Latin American, have certain height and weight, and many other characteristics. Each characteristic defines a different population. For which of these populations should we apply the base rate?

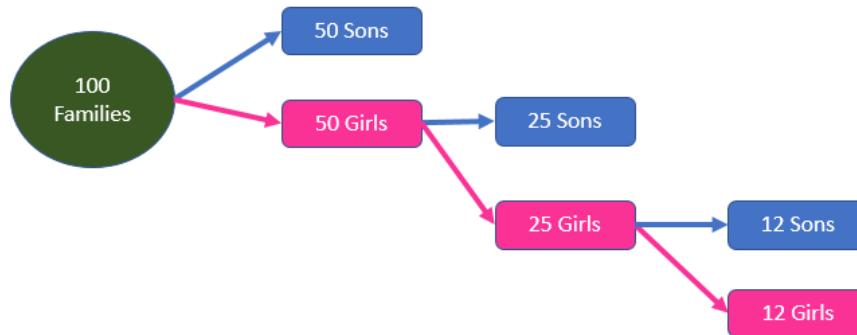
Another way to think about this issue is about how to USE this probability. If one is analyzing costs and benefits from screening, this is a useful statistic, because it tells of the percentages which will fall into various categories. Once the screening has been done on a particular woman and it comes out positive, her probabilities of having cancer are much higher than the base rate, regardless of the number of false positives. The base rate is just one source of information regarding the probability of cancer. We should look to other, stronger, more conclusive, sources of evidence, which will be available for a specific person. After the outcome of a random event has occurred, there will normally be many more sources of knowledge regarding this outcome than just the prior probability of occurrence. The base rate should be ignored, and attempts should be made to determine why screening gave positive result on this specific person.

Son Preference

The map shows prevalence of son-preference, demonstrated by high proportion of male infants in comparison with female infants. We will set up an artificial probability model to see if “stopping rules” can create an imbalance among the sexes, as many believe. According to this artificial model, the goal of all families in a certain country is to have ONE son. Thus, they continue to have children until they achieve this goal. They stop after the first son is born.



A time-branching probability model for this situation can be given as follows:



Among 100 imagined families, exactly 50 of them have sons on the first try, while 50 have daughters. Those with daughters go on to the second child. For the second child, 25 families have sons, while 25 have daughters. Those with two daughters go on to the third child. It is easy to see that at each stage the numbers of sons and daughters is balanced. Regardless of when families choose to stop, they cannot influence the ratio of girls to boys in the general population.

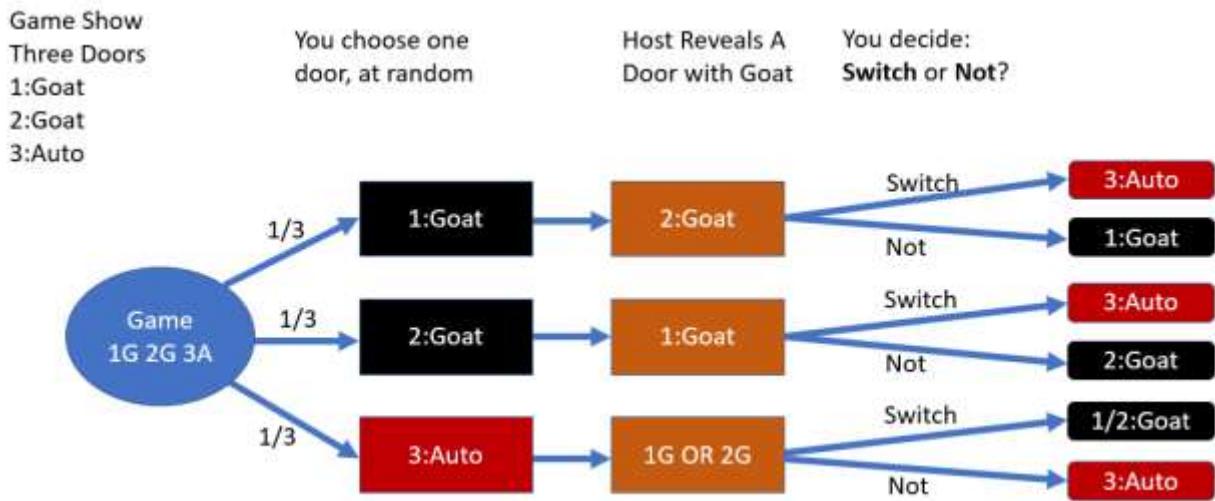
From this analysis, we see that stopping rules cannot explain population imbalances. Thus we must look beyond to explain the “Missing Girls” in son-preference cultures. Two of the obvious ones are selective abortions and inequitable health care. If female infants receive poorer nutrition and healthcare in comparison with males, than this will create the observed gender imbalance.

Marylin Vos Savant & The Goats:

As our last example, we analyze a game-show based probability puzzle, where the intuition of top mathematicians & physicists fails to match the official, widely accepted answer. The Wikipedia article details of the enormous amount of controversy surrounding [The Monty Hall Problem](#). The problem lies, Once again, failure to differentiate between pre-event

probability and post-event confidence is the source of the confusion. We first set out the problem, together with the standard analysis.

You are the contestant in a game show “Let’s Make A Deal”. You have to choose between three doors, one of which contains the grand prize of an automobile, while the other two have goats behind them. After you choose a door, but before revealing what is behind the door, the host offers you an option. He opens one of the doors you have not chosen, and shows you that it has a goat behind it. He then offers you the option to switch your choice. This time-branch sequence for the game is drawn below:



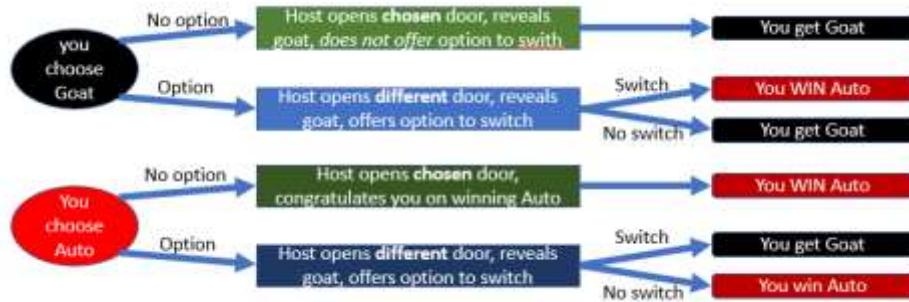
Standard analysis of this game comes to the (wrong) conclusion that you should switch. This is based on the pre-choice calculations, which can be made as follows. In two cases you have chosen a goat. In both these case, switching will give you the Automobile. In the third case, you have chosen the automobile, and switching will give you the Goat. Thus, pre-choice calculations show a 2/3 probability of winning by switching. If you do not switch, you have your original 1/3 chance of having chosen the Automobile. This seems to show that it is favorable to switch.

This analysis is wrong because AFTER you have made the choice, probabilities have been extinguished. The host KNOWS whether you have chosen a goat or an auto. We must therefore analyze

Standing at time t=0, BEFORE making choice, probability 1/3 of selecting the right door.

At time t=1, choice has been made. Probability has been extinguished. Host KNOWS whether or not you have made correct choice. Any action on part of the host must be interpreted as a signal to you regarding this knowledge. How we should interpret the signal – his showing you the goat and giving you the option to switch – depends on the intentions of the host. The game should be analyzed as a DETERMINISTIC game of STRATEGY, with asymmetric

information. We need to use psychology, not probability, to analyze the game. Here is alternative model of the game, starting from the two deterministic possible outcomes of the choice you have already made:



The host has two possible moves. He can either just open the door and show you what you chose. Or, he can offer you an option to switch, after showing you a goat. The resulting game tree is listed above. If the host is playing AGAINST you, then it is clear that he should show you the goat if you have chosen it. In case you have chosen the auto, he can offer you an option, in order to psych you into losing by switching.

We can see that this is not a question about probability by noting that analysis depends strongly on the intentions of the host. If the host is adversarial, trying to win against you, then you should not switch. It is clear that any message he sends you cannot be to your advantage. If you can reach agreement with the host to split the prize, so that you can count on his good will, then the host can ensure that you win by offering you a choice only when you choose the goat.

What is the right model for host behavior? We can look at the empirical evidence. In the popular movie Slumdog Millionaire, the host of the gameshow builds up trust in order to deceive the contestant by giving him a false clue for the million-dollar final question. The original host Monty Hall demonstrated how he could win 8 out of 8 sample games against contestants by psyching them into switching when they chose correctly, and not offering them an option when they chose the goat. The standard analysis comes to the wrong conclusion because it uses pre-choice probability measure for the post-choice situation, where the outcome is determined and known to the host.

Concluding Remarks

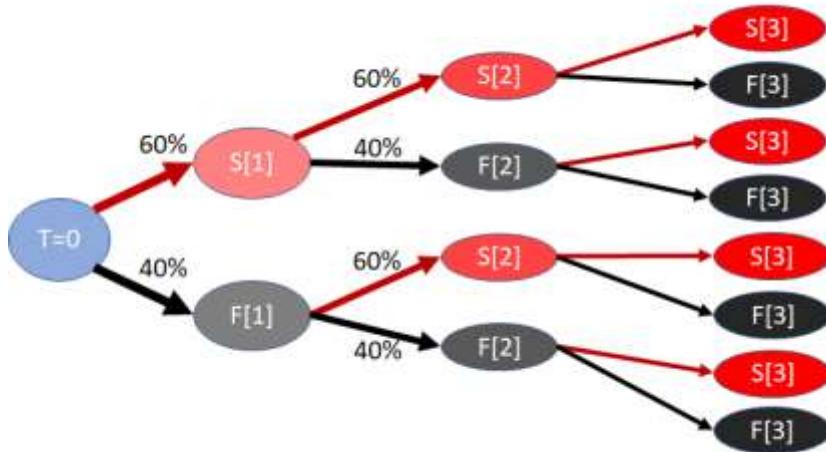
The goal of this lecture was to demonstrate how making and using Time-Branching Probability Models can lead to clarity in MANY confusing probability calculations. HUGE numbers of confusions arise about probability because standard theory and notation IGNORES the temporal nature of probability. The key fact that $P(X=H)=50\%$ holds only BEFORE coin is flipped and NOT afterwards is not mentioned in textbooks as an essential feature of probability. Time Branching Probabilities create a picture of a world radically different from the Newtonian deterministic world. At each moment, each choice we make creates a branch to a new world, while at the same time removing large numbers of other possibilities. All the branches which were possible before the choice now pass into the shadowy realm of what might have been.

[RSIA08C Examples TBPM.docx](#)[RSIA08C Applic of TBPM.pptx](#)

7D Binomial Probabilities

The Binomial is the simplest possible probability model. At T=0, many different outcomes are possible. Group outcomes into TWO sets – call one set SUCCESS, rest FAILURE. The combined probability of all outcomes which make up the event S is p, while the probability of F is 1-p. Now suppose that this model repeats at every later outcome. After each outcome, at the next time period, there are two branches, to S and F. Furthermore, the probability of S and F remain fixed as we progress in time, and is also fixed on all branches at any particular point in time. This can be seen as a sequence of independent trials, where on each trial the probability of success is p. We can also model this as a sequence of coin flips which comes out Heads (Success) or Tails (Failure). Note the coin is BIASED and P(H)=p, which is not necessarily 50%. Our goal in this lecture is to analyze this probability model.

The Case of Three Trials:



At T=0, there are two possible outcomes, S & F, with probabilities 60% and 40%. At each future node, the same branches repeat, with the same probabilities. Since branches double at each point in time, the total number of possible outcomes for T=3 is $2 \times 2 \times 2 = 8$. For each outcome, the probability is the product of the branches leading to it. For example, for FSS, the probabilities are $40\% \times 60\% \times 60\%$, the product of the probability for the first F, followed by the probabilities for the 2 S's. Since the probabilities do not depend on time, we can re-arrange the sequence without changing the probability. Next, we consider generalizing this to N time periods, with N trials.

Independent Repeats: Bernoulli RVs

To go from $T=3$ to $T=N$, note that at each future time, number of POSSIBLE outcomes doubles from the previous one. Thus, the total number of possible outcomes at $T=N$, counting from $T=0$, is: $2 \times 2 \times \dots \times 2 = 2^N$. One outcome can be written as a sequence of S's and F's where the total number of elements in the sequence is N . Because of independence, sequencing does not matter for the probability. Any sequence of length N , O = SSFSF ... FFSFF has probability depending on number of S's (K) and number of F's ($N-K$). This probability is: $P(O) = p^K \times (1-p)^{N-K}$. With these basic ideas in place, we can now define the central concept of this lecture:

Definition of Binomial Random Variable: X is a Binomial Random Variable: $X \sim Bi(N,p)$ if X is the COUNT of the number of successes in N independent trials such that the probability of success in each trial is p.

From the definition, it is obvious that X takes values between 0 and N – within N trials there can be a maximum of N successes, and a minimum of 0. For each trial with K successes and $N-K$ failures, the probability is $p^K \times (1-p)^{N-K}$. In order to compute the probability of K successes, we need to count the number of outcomes which have K successes and $N-K$ failures. This leads us to the famous:

N choose K formula: First we develop the formula in a special case where $N=10$ and $K=4$. Given a sequence of 10 Time periods, how can we choose 4 of these time periods to put Heads (success) into? The rest of the periods will be Tails (Failure). Note that for a sequence like HTHTT HTTTH with 4 H, 6 T, the probability will be $p^4 (1-p)^6$ since the success probability of Heads is $p=0.6$ while the probability of failure (Tails is 40%). But HOW many sequences are there with 4 Heads? Here is a different example: HHTTT TTTHH. The formula for N choose K answers the question of “How many ways can we put exactly K Heads in N Trials?”. This number is called N choose K and denoted $C(N,K)$:

Detailed derivation of this formula is given at the end of this lecture, in an appendix. For now, we take the formula as given, and proceed to use it to derive the properties of Binomial Random Variables.

Binomial Probabilities

If $X \sim Bi(N,p)$ then X is the number of Successes in a sequence of N Trials, where the probability of success in each trial is p, and all the trials are independent of each other. X is a random variable. The number of successes in N trials can range from 0 to N, so X can take values from 0 to N. We want to compute $P(X=K)$. For each sequence of trials with K Successes and $N-K$ Failures, the probability is $P = p^K \times (1-p)^{N-K}$. To get $P(X=K)$, we need to count the NUMBER of sequences with K successes. This number is given by $C(N,K)$, which measures the number of ways to put exactly K Successes into a sequence of N trials. We multiply the number of outcomes with K Successes with the probability of each outcome to get the desired Binomial probability.

Properties of Binomials

We have skipped over the technical details in deriving the formula for Binomial Probabilities, since intuitive understanding of properties of this distribution is far more important

than ability to do mathematical derivations. The calculations of the probabilities can be left to the computer, but the computer cannot understand what they mean, or how they are used in real world applications.

Practical Probability

In the real world, we use probabilities to assess the chances of different possible futures, and to take decisions in the light of our understanding. Some basic translations of technical and quantitative probabilities into qualitative aspects of our beliefs about the future can be listed as follows:

1. If $P(A) > P(B)$, then A is more likely to happen than B.
2. If $P(A) \geq 50\%$, then A is more likely to happen than not.
3. If $90\% \leq P(A) \leq 99\%$, A is very likely to happen.
4. If $99\% < P(A)$, A is almost a sure shot

ALL of these are QUALITATIVE terms. Numbers have been attached as ROUGH GUIDES. We use intuitive and qualitative judgments regarding probabilities to make decisions. This is in contrast with the dominant view: Rational Decision Making theory asserts that we assess numerical probabilities for all outcomes, and evaluate all outcomes in order to arrive at decisions. This theory led to massive failure in the Global Financial Crisis. We are trying to develop an alternative in this course.

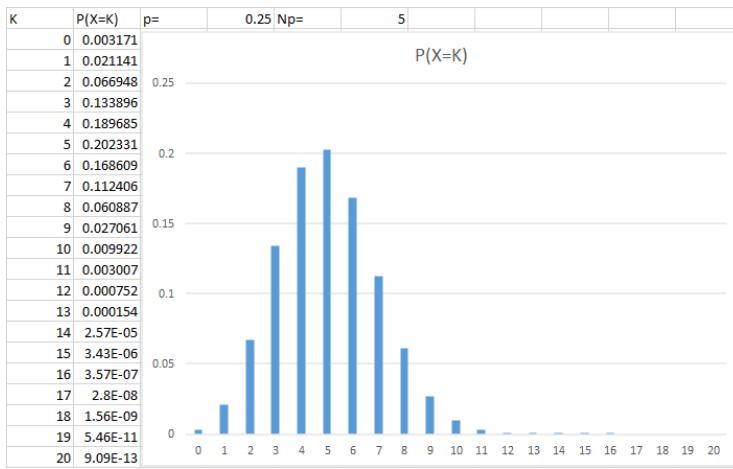
GOAL: Our analysis of Binomial probabilities will teach us to recognize behavior of Binomial Random Variables. We will create Binomial models for various types of real world events. Then we will match properties of Binomial models to properties of real world data. If match is poor, Binomial model does not fit. This will lead to some conclusions – (rejection of null hypothesis of GOOD MATCH of MODEL). Rejection of model will lead us to search for alternative, better, models. If match is good, we TENTATIVELY accept null hypothesis of match. This allows us to interpret and explain the data in a deeper way than just the mere observations. Tentative acceptance of the null hypothesis of a good match may change if later data reveals mismatch.

How do we match probability models to data?

Probability models are time-branching trees, only one path out of zillions of possibilities is actually realized. This makes it impossible to directly observe truth of any such model. All such models create structures of unrealized and unobservable possibilities. Indirect confirmation of the model is achieved by looking at the predictions of the model. In particular, models tell us “What is VERY LIKELY to happen (99% confidence)?” If observations fall within this range, then model is judged to be acceptable. If observations fall outside – what model predicts is likely to happen DOES NOT happen – then model is unreliable.

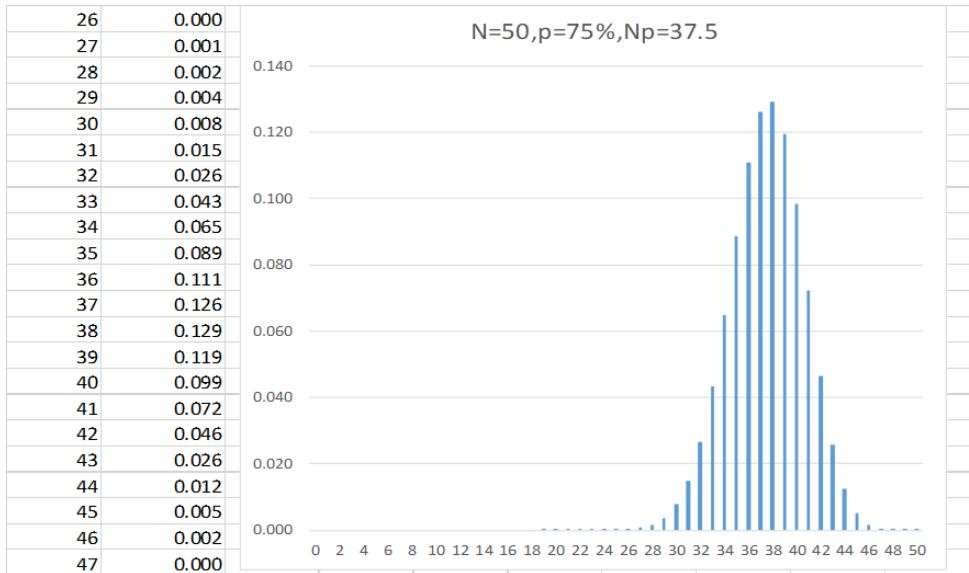
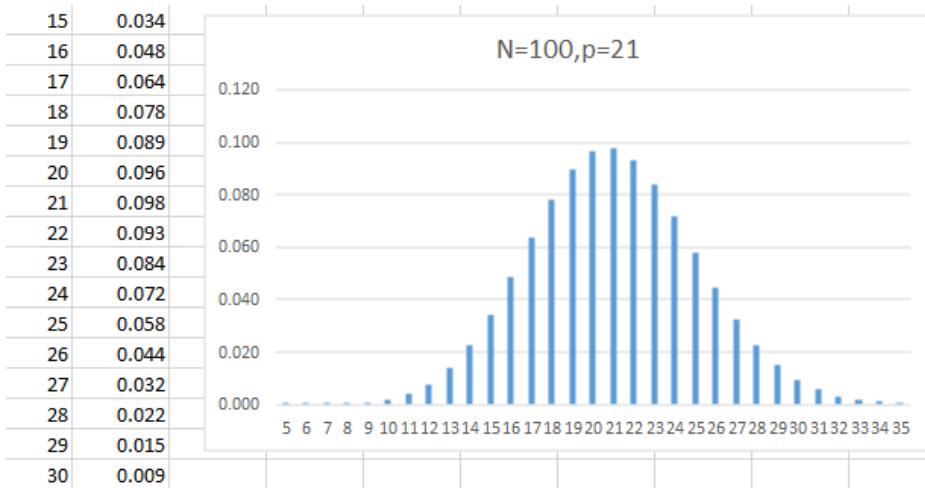
NEXT GOAL: In line with this purpose, we will find out what is likely to happen with Binomial models. A promising place to start is to look at the MODAL value: this is the most

likely outcome for the number of successes in a sequence of N trial. It turns out that this conforms to what we would intuitively expect. If Np is an integer, than the Modal value is Np . If Np is not an integer, then the modal value is one of the two integers between which Np lies. It is easy to explain why. The probability of success is exactly the proportion of successes that we would expect to see in N trials. So the value of K such that K/N is closest to the success probability p is the most likely value of X when $X \sim Bi(N,p)$. We will now compute probabilities associated with such variables to understand how Binomial probabilities behave.



Case of $Bi(20,0.25)$: X is Binomial: $N=20$, $p=0.25$; Modal Value: $Np=5$. Given 20 trials where the success probability is 25% on each trial, the most likely number of successes is 5. The Binomial probabilities and their graph is given below. We can see from the table produced by EXCEL that $P(X=5)=0.202$. This probability of 20.2% is highest at MODE and declines on both sides. $P(X=4)$ is about 19% while $P(X=6)$ is about 17%. This is a UNIMODAL distribution, it has only one peak value. The interval from 1 to 9 [1,9] is range of Central Values. Values of 0 and those above 10 are highly unlikely, have probabilities less than 1 in 1000. This means that if we create this model for a particular real world situation, and observe outcomes larger than 10, we can reject the model. What the model predicts as extremely unlikely was observed, leading to the conclusion that the model is probably wrong.

Case of $Bi(50,0.75)$: In this case, $Np=37.5$. When Np is not an integer, the mode can be either of two nearest integer values. In this case, the MODE=38, and $P(X=38)=12.9\%$. As usual, the distribution is unimodal. The Central Range of [27,46] contains all points with probability greater than 1 in 1000. We should reject the model if the observed number of successes lie outside this range.

**Binomial with $N=100$ and $p=21\%$** 

Case of Bi(100,0.21): With 21% success probability, we expect to see $Np=21$ successes in 100 Trials. The Mode is 21, and $P(X=21)=9.8\%$, which is the highest probability among all possible outcomes of X . As usual, this is a Unimodal distribution, and the central values: [9,34]. Outside this range, probabilities of outcomes are less than 1 in 1000. If we use this probability model for a real world situation, and observe that the number of successes is less than 9 or greater than 34, then we should reject this model.

Case of Bi(500,18%): Our final example is that of a Binomial Random variable with $N=500$ and $p=18\%$. In this case, $Np=90$ is the MODE and $P(X=90)=4.6\%$ is the highest probability among all the outcomes. Also the range of central values is within 30 units of the MODE 90. That is, outside the range [60,120], all outcomes have extremely small probabilities with $p<0.0001$. Obviously, we should reject this model if observations are out of this range.

$$X \sim Bi(N, p) \Rightarrow P(X = K) = \frac{N!}{K!(N-K)!} p^K (1-p)^{N-K}$$

Lessons from these graphs of Binomial Probabilities

It makes sense that the MODAL value is Np and that this has the highest probability. HOWEVER, on initial encounter, it seems surprising that the probability of MODE goes to 0% as N increases. The reason is that as N increases, probability SPREADS out over a large number of points. The central range EXPANDS in terms of absolute size. Basically our probability model predicts that the number of successes we observe will be within the central values. In the models examined above, the prediction of the probability model can be summarized as follows:

$$N=20 \quad p=0.25 \Rightarrow [1,9] \quad \text{range } 8/20 = 40\%$$

$$N=50 \quad p=0.75 \Rightarrow [27,46] \quad \text{range } 19/50 = 38\%$$

$$N=100 \quad p=0.21 \Rightarrow [9,34] \quad \text{range } 25/100 = 25\%$$

$$N=500 \quad p=0.18 \Rightarrow [60,120] \quad \text{range } 60/500 = 12\%$$

Note that the SIZE of the interval of central values GROWS: 8, 19, 25, 60. However, when compared to the full range, the size of this interval SHRINKS in percentage terms: 40%, 38%, 25%, 12%. The accuracy of prediction of the probability model in percentage terms increases.

Key asymptotic results (large N)

General principles about what happens when N is large come from sophisticated mathematics which can be used to approximate what will happen in this (asymptotic) case. The key results, which have practical value are the following.

Law of Large Numbers: X is the number of successes in N trials, where the probability of success in each trial is p . The ratio X/N is the percentage of successes in N trials. The Law of Large Numbers says that X/N will be close to p when N is large. That is, the OBSERVED proportion of successes will be close to the theoretical probability of success p specified by our probability model.

Central Limit Theorem: A more precise prediction is made by the Central Limit Theorem. First, define the Standard Deviation (SD) by the following formula: . Then the difference between X/N and p will be within the following limits with high probability:

$$-2SD \leq (p-X/N) \leq 2SD \text{ holds with 95\% probability for large } N$$

$$-3SD \leq (p-X/N) \leq 3SD \text{ holds with 99.7\% probability for large } N$$

These theoretical results give us high confidence intervals for assessing match of probability model with observations. In a later lecture, we will show how to apply these results to assess probability models, and how these results give us deeper insights into observational data.

LINKS

A New Definition of Probability: <http://bit.ly/rsia08a>

Time-Branching Probabilities: Rules of Calculation: <http://bit.ly/rsia08b>

Applications: Time-Branching Probability Models: <http://bit.ly/rsia08c>

The Binomial Distribution: <http://bit.ly/rsia08d>

Online Course: Real Statistics: An Islamic Approach <http://bit.ly/dsia786>

APPENDIX: The N Choose K Formula.

How many ways to put K heads into a sequence of N trials? It is easier to understand solutions to abstract algebraic problems like this by first solving them for particular values of K and N. This makes it easier to understand intuitively the logic behind the answer. So we will first solve this problem for the special case that N=10 and K=4. How many ways can we put exactly four 1's into a sequence of 10 slots, each of which can contain 0 or 1?

Special Case: N=10 and K=4

To solve this problem, it is useful to break it down into two steps. The first step is to solve a simpler problem.

FIRST STEP: Instead of putting four 1's, we put four DIFFERENT items into four slots within a sequence of 10 slots. Let us name the four items A,B,C,D. Now we have four different objects, and we want to count how many ways we can put these into 10 slots. Now the basic counting formula provides a simple answer. The first object A can be put into any of 10 positions. The second object B can be put into any of the 9 remaining positions. The third object C can be put into any one of the 8 remaining positions. The fourth object D can be put into any of the remaining 7 positions. So the answer to this question is $10 \times 9 \times 8 \times 7$. In mathematical notation, $N!$ is the product of all integers from 1 to N, so $10! = 10 \times 9 \times 8 \times \dots \times 1$. If we want to go from 10 to 7, then we can eliminate the factors 6, 5, 4, 3, 2 by dividing by $6!$. Using this notation, we can say that the number of ways to put 4 different objects into a sequence of 10 slots is $10! / 6!$

SECOND STEP: How much overcount? The above formula is not the answer to our original question because it has an overcount. To see why consider the following sequences: C0A0B0D0000, A0C0B0D0000, D0C0A0B0000. These are all sequence of 10, with 6 0's and four places with A, B, C, and D in them. If we replace A,B,C,D by 1, then all three will be the SAME sequence: 10101010000. But our first step formula counts all three of these as separate sequences. So the question is, how much overcounting do we do when we count different arrangements of A,B,C,D as separate sequences? Let us consider just one sequence 1111000000, where all four ones occur in the initial four positions. How many times is this sequence counted when we replace four 1's by four different objects A,B,C,D?

All the ways we can put ABCD into the first four slots are equivalent when we replace ABCD by 1111. The first A can be put into any of 4 positions. The second object B can be put

into any of the remaining 3 positions. The third object C can be put into any of the remain 2 positions. The last object D will have only one remaining position into which it can go. So the total number of ways we can put ABCD into the first four positions is $4! = 4 \times 3 \times 2 \times 1$.

This same reasoning holds for ANY placement of the four 1's into the sequence of ten 0's and 1's. Every such sequence will be overcounted $4!$ times. Thus we get the answer to the original question by dividing the answer of the first step by $4!$. The number of ways to put exactly 4 1's into a sequence of 10 0's and 1's is $10! / [6! \times 4!]$.

The General Case for any N and K

We can now just repeat the same reasoning to get the general answer. Suppose we have a sequence of N 0's and 1's. We want to count how many ways we can put exactly K 1's into this sequence.

FIRST STEP is to replace the identical objects "1" by distinct objects; let us name them 1,2,3,...,K. The first object can be placed into any one of N positions. The second one can be put into N-1. Continuing in this way, the K-th object can put into any one of N-(K-1) positions. Multiplying these choices, the answer at the first step is: $N \times (N-1) \times \dots \times (N-K+1) = N! / (N-K)!$

SECOND STEP: How much overcounting do we do when we replace identical 1's by different objects? The sequence of numbers 1,2,...,K can be re-arranged in $K!$ different ways. We can put the 1 into any one of K positions, the 2 into any of the remaining K-1 positions, and so one. All of the $K!$ ways are identical when we replace the different objects by the same object "1". This means that there is an overcount of $K!$ in the first step. Thus the solution to the original problem divides the answer of the first step by $K!$ arriving at the N choose K formula: $N! / [K! \times (N-K)!]$

Binomial Probabilities.

We can now state the final result of all of these calculations as follows. Suppose we have a sequence of N independent random events which have two possible outcomes named "1" and "0". Suppose the probability of "1" is p for each of the N events. Let X by the random variable which counts the number of "1"'s which occur in this random trail. This situation is described by writing . This random variable X can take any integer value from 0 to N. The probabilities of each of these possible outcomes, PRIOR to the occurrence of the N events, is given by the following formula:

This formula is obtained by noting that when there are K 1's and N-K 0's, the probability of this outcome is $p^K (1-p)^{N-K}$. The number of outcomes which have K 1's and N-K 0's is N choose K which is $N! / [K! \times (N-K)!]$. Adding up the probability $p^K (1-p)^{N-K}$ for all of these outcomes leads to the formula given above.

7E Random Sampling

Surveys of Large Populations

To a much greater extent than our materialistic theories recognize, ideas rule the world. It is often important to learn how people feel and what they think about various issues. Surveys of opinions within a population provide one important way to collect such information. In this lecture, we will concentrate on the case of voting behavior. How can we learn about the opinions of the population of the country regarding various presidential candidates? The straightforward method of asking everyone is impossibly expensive and time consuming. We cannot set up a survey of millions of people. This leads to the necessity of *samples*. Instead of asking the entire population, we could ask a small number of people, and hope that this smaller sample will be representative of the entire population.

To make the issue sharp and concrete, we focus on the pollsters problem: “How to determine what will happen in the elections?”. We will simplify matters by taking the best case scenario for pollster. Suppose that everyone will vote, and everyone is decided on who they will vote for. Even in this best case scenario, it is impossible to count EVERYONE. We must take a SAMPLE, and use the sample to infer population results

Pollster Failures & Learning from Experience

We can HOPE that our sample will provide us information about the general population. However practical experience shows that this hope is often unjustified. My personal experience at MIT testifies to the dramatic failure which is possibl. In the election year 1972, Nixon was running against Brown. Everyone I know was strongly in favor of Brown. So, on the basis of the sample of the people I knew, I was confident that Brown would win the election. In fact, Nixon won a landslide victory, winning 49 states, while Brown won only one state, Massachusetts! I happened to be in the wrong state. Even if I had taken a systematic and large sample in the whole state, I would have come up with the wrong prediction for the election.

There are many examples of disastrously wrong forecasts by pollsters. Freedman’s textbook on Statistics provides a case study of the Literary Digest Poll in 1936. The LD sample was largest ever: 2.4 Million People. But it gave a disastrously wrong result – LD went out of business soon afterwards! The most recent case was in 2016, when Donald Trump ran against Hilary Clinton. The vast majority of pollsters predicted a win for Clinton, but Trump won the election. Because it is an excellent example of the power of the Binomial Model, we will study the question of “How do pollsters like Gallup predict election outcomes?”

Problem of REPRESENTATION

About 155 Million people voted in USA 2020 for Trump vs. Biden. How to predict outcome of elections? Biden won 81,283,098 votes; Trump won 74,222,958 vote. It is obviously impossible, or, very costly and difficult, to POLL All 155 Million. Pollsters like Gallup use only about 3000 people! Is it possible to choose such a small sample in a way which is

REPRESENTATIVE of the whole population of 155 Million? To be more precise, consider the proportion of Biden voters in the sample of 3000 people who were surveyed. How does this PROPORTION within sample reflect the PROPORTION in Population? If we pick just any 3000 people, it would be possible for all of them to be Biden voters, and also for all of them to be Trump voters. So the problem is: HOW to choose a sample which accurately reflects the whole population?

Random Sampling

We have spent a lot of time criticizing Sir Ronald Fisher, the founder of modern statistics. Here, we must acknowledge the genius of Fisher for developing the concept of random sampling, as a way getting samples which represent the population. Intuitively, we think that there must be some systematic method for choosing samples which would work better than random, haphazard, choice. However, Fisher showed that the opposite is true. While no systematic method works well, it is an amazing fact that if we pick a sample “AT RANDOM” – it will reflect the population. Here are three counter-intuitive facts about random sampling:

1. Proportions of voters in random samples reflect population proportions
2. All SYSTEMATIC methods of picking samples FAIL !!
3. Size of population does not matter, accuracy of sample depends on size of random sample we take (3000 is enough for 150 Million!)

Before proceeding, it is essential to clarify the two senses of the word RANDOM. In standard English language sense, any method without any systematic pattern is random. However, this is very different from the way statisticians use the same word.

TECHNICAL STATISTICAL SENSE: To pick one person at random from a population of 1000, every one must have an EQUAL CHANCE of being selected. This is VERY DIFFICULT to achieve in practice.

These two senses are often CONFUSED. To achieve greater clarity, consider “How to achieve technical randomness?” Suppose we want to choose one person “at random” from among 1000. First, we must make a list of 1000 names. Then we must use a randomization mechanism to pick one of these names. Randomization mechanisms can be based on pseudo-random generators on computer, with real-time seed selection. We need to be able to show that ALL people on the list had an equal chance of being selected. This DEMONSTRATION of technical randomness is always done within a probability model of the event. It is NEVER possible in reality, because we can never show what MIGHT have happened. The demonstration involves showing that even though X was chosen, all other members of the population ALSO had equal chance of being chosen. This can never be proven/demonstrated EMPIRICALLY. What we can do with our random number generators is to show that they go through to generate all the numbers within 1 to 1000 with roughly equal frequency, and without any predictable patterns. This is the closest we can come to demonstrating that a particular choice is random. This means that many aspects of our probability models are unobservables.

Properties of Random Samples

We will now develop properties of random samples within the framework of the Binomial Distribution developed in the previous lecture 8D. We consider a sequence of RANDOM picks from a population of 155 Million voters, with 81M for Biden and 74M for Trump. All voters have equal chance of being selected. Let SUCCESS or “1” be Biden voter, and FAILURE or “0” be Trump Voter. Our first random choice S1 has Binomial probabilities p and 1-p for being “1” or “0”. That is, $P(S1=1) = 81/155$, $P(S1=0)=74/155$. Next, consider a sequence of random choices: S1, S2, ..., S(N). This is a random sample of size N. It is a sequence of 1’s and 0’s – that is, Biden voters or Trump voters. Let X be the total number of Biden voters in the sample. Then X is Binomial with N trials and 81/155 probability of success. We can now apply Binomial theory results. Note that probability results are PRIOR to taking the sample. The proportion of Biden voters within the sample is X/N which we will observe after we take the sample. Within the population, the actual vote was 81/155 for Biden. This number is NOT known at the time we are doing sampling – indeed, the GOAL of sampling is to LEARN what this number is. This number is the success probability p for each of the Binomial trials – p is the (unknown) chance of picking a Biden voter in our random sample. The relationship between the sample proportion X/N and the population proportion p can be expressed in the following way.

Large Sample Confidence Intervals

Given that X (the number of Biden voters in our random sample of size N) is Binomial with success probability p, we can compute the probability distribution of X. The modal value of X/N is p, and the interval around this value which contains MOST of the probability can be computed with high accuracy for large values of N using large sample (asymptotic) theory. The key results from this asymptotic theory can be expressed in terms of the SD (Standard Deviation) as follows. First, we must compute the SD:

$$SD = \frac{\sqrt{p(1-p)}}{\sqrt{N}} = \frac{\sqrt{52.2\% \times 47.8\%}}{\sqrt{3000}} = \frac{0.4995}{54.77} = 0.0091$$

It is important to note that we CANNOT actually compute the SD above, because we do not know “p” – the true proportion of Biden voters in the entire population. This is the number that we are trying to find out using our survey. The observed proportion of Biden voters in our sample is X/N, and that is our best estimate of p. We can use this estimate to get an approximate value of p, and that allows us to get an approximate value of SD. The SD is not very sensitive to this approximation in large samples. For example, in the above computation, the true value of p=52.2% leads to 0.4995 in the numerator. If we replace this by 50%, we would get 0.5000 which is only different by 0.0005. Small changes in p have very little effect on the numerator of the SD.

Once we have the SD, we can construct the following large sample confidence intervals for p, the true proportion of Biden voters in the population as a whole. The standard confidence

interval creates an interval of 2SD's and has 95% confidence: $-2SD \leq (X/N) - p \leq +2SD$. This is equivalent to $|(X/N)-p| \leq 2 SD = 1.8\%$. That is, with 95% confidence, the sample proportion will be within 1.8% of the true proportion p of Biden voters in the full population. We can get a higher degree of confidence by expanding the interval to 3SD. In this case, we can say that $|(X/N)-p| \leq 2.7\%$ with 99.7% confidence. These large sample confidence intervals tell us how accurately the proportion of Biden voters in our random sample approximate the true proportion which we wish to forecast. We can say that, BEFORE carrying out the poll, the probability is 95% that the sample will give us a forecast within 1.8% of the true percentage vote. Also, with probability 99.7%, the sample proportion will be within 2.7% of the true percentage of Biden voters within the full population of voters. What is surprising is that these statements about how accurately the sample predicts the population depends ONLY on size of random sample, and not on the size of the Population. A second complication is that, after the poll, these statements turn into confidence statements. Lets say you make a prediction that the proportion of voters for Biden will lie within 1.8% of 51% (which is proportion of Biden voters you get in your random sample). Before taking the sample, this is true with 95% probability. AFTER the sample is taken, the statement is either TRUE or FALSE. There is no longer any probability associated with it. The true proportion either falls into the interval or not. Probability is either 100% or 0%, not 99.7 or 95 – but you have no knowledge of the truth. SUBJECTIVE probability might match confidence levels, but could be adjusted on basis of other knowledge.

How SD behaves as function of N

As we have seen, the accuracy of a sample depends only on the sample size N , and on the value of p . To get an idea of how much accuracy we can achieve using random samples, the following table lists the SD, with provides a 95% confidence interval, for various values of N and p – note that p and $1-p$ are equivalent for SD:

| (N,p) | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|-------|-------|-------|-------|-------|-------|
| 100 | 0.030 | 0.040 | 0.046 | 0.049 | 0.050 |
| 500 | 0.013 | 0.018 | 0.020 | 0.022 | 0.022 |
| 1000 | 0.009 | 0.013 | 0.014 | 0.015 | 0.016 |
| 5000 | 0.004 | 0.006 | 0.006 | 0.007 | 0.007 |
| 10000 | 0.003 | 0.004 | 0.005 | 0.005 | 0.005 |
| 50000 | 0.001 | 0.002 | 0.002 | 0.002 | 0.002 |

Note that the smallest amount of accuracy occurs at $p=50\%$. Even with a small sample of 100, a 95% confidence accuracy of 10% (=2SD) can be achieved. At sample size of 5000, with 95% confidence, the error of the sample estimate will be 1.4%. At 50,000, the error is only 0.4% with 95% confidence. Thus we see that when the voting percentages of the two candidates are very different, even a small sample can give correct forecasts. However if race is CLOSE, we

need much larger samples to get desired accuracy. The size of the population is irrelevant for this purpose.

Empirical Versus Theoretical Accuracy

We have now discussed the theoretical properties of random samples based on the Binomial model. How does random sampling behave in real life? How well does theory match practice? The table displays the errors made by the Gallup poll, which uses random sampling to produce forecasts of the popular vote. It is easily seen that the errors are much larger than theoretical predictions based on the Binomial model. HOWEVER, random sampling works well at producing accurate results most of the time. Over time, size of sample being taken has gone down, and accuracy has gone up. Partly, this is due to improvements in the “randomness” of sampling. But it is also due to learning from experience. When errors are made, pollsters investigate the sources of error, and attempt to find ways to overcome them. We next discuss the sources of errors in opinion polls like this one.

Table 4. The Gallup Poll record in presidential elections after 1948.

| Year | Sample size | Winning candidate | Gallup Poll prediction | Election result | Error |
|------|-------------|-------------------|------------------------|-----------------|-----------|
| 1952 | 5,385 | Eisenhower | 51% | 55.1% | 4.1% |
| 1956 | 8,144 | Eisenhower | 59.5% | 57.4% | 2.1% |
| 1960 | 8,015 | Kennedy | 51% | 49.7% | 1.3% |
| 1964 | 6,625 | Johnson | 64% | 61.1% | 2.9% |
| 1968 | 4,414 | Nixon | 43% | 43.4% | 0.4 of 1% |
| 1972 | 3,689 | Nixon | 62% | 60.7% | 1.3% |
| 1976 | 3,439 | Carter | 48% | 50.1% | 2.1% |
| 1980 | 3,500 | Reagan | 47% | 50.7% | 3.7% |
| 1984 | 3,456 | Reagan | 59% | 58.8% | 0.2 of 1% |
| 1988 | 4,089 | Bush | 56% | 53.4% | 2.6% |
| 1992 | 2,019 | Clinton | 49% | 43.0% | 6.0% |
| 1996 | 2,895 | Clinton | 52% | 49.2% | 2.8% |
| 2000 | 3,571 | Bush | 48% | 47.9% | 0.1 of 1% |
| 2004 | 2,014 | Bush | 49% | 50.6% | 1.6% |

Note: The percentages are of the popular vote. The error is the absolute difference “predicted – actual.”

Source: The Gallup Poll (American Institute of Public Opinion) for predictions; *Statistical Abstract*, 2006, Table 384 for actuals.

Reasons for mismatch between theory and practice

We list briefly some of the main difficulties encountered in opinion surveys based on random sampling.

1. Before the recent dramatic increases in information processing capabilities, it was not possible to get a list of all 150 million registered voters. In such situations, using

something called multi-stage cluster sampling, it is possible to produce a “probability” sample. A probability sample has the characteristic that we can COMPUTE the probability of selection for each individual, but they are not exactly the same across the entire population. This, and other potential errors in the choosing the sample, can lead to higher errors than the theoretical formula based on SD.

2. Selection Bias: When the sample chosen for the survey differs systematically from the population on essential characteristics, this is called selection bias. Theoretically, random sampling eliminates selection bias, which was present in earlier methods used by pollsters. For example, see discussion of Quota Sampling in Freedman excerpt.
3. Non-response Bias: After selecting a sample, there is the problem of getting them to answer the questions asked. This requires being able to contact the person, and also of getting them to cooperate in giving you an answer. If people who are easy to contact and willing to respond are different in systematic ways from those who are harder to contact and less willing to respond, this will create a bias in the survey. The bias will be due to people who do not respond.
4. Response Bias: Sometimes the way the question is asked, or the appearance and attitude of the questioner, influences the responder to answer the question in certain ways. This can create a bias. For example, it is known that people like to appear conscientious. If you ask whether or not they plan to go to the poll, they are likely to say “yes” to make a good impression, even if they do not plan to go. This is called the response bias.
5. Because the voting will take place in the future, pollsters face the problem of deciding “who will actually GO to the polls?” There is no point in sampling opinions of those who will not cast a vote. Similarly, there can be people who are undecided; how they will make up their minds in the future is not easy to forecast. In addition, there are people who SWITCH their vote between poll and election.

For all these reasons, elections are hard to forecast. Also, the forecast errors do not match the theoretical errors of the Binomial Distribution. Freedman provides many real-world examples of failures of pollsters. Many more have occurred since then, with 2016 Trump versus Clinton being a famous example. Forecasting the future is inherently difficult. This shows us that the Keynes-Knight model of RADICAL UNCERTAINTY provides us with a better guide than the standard rational choice model.

Concluding Remarks

To wrap up this lecture, we note that Random Sampling is one of the MOST IMPORTANT applications of probability to real world. Surveys without random samples are not reliable. Samples-of-convenience can fail to be representative in very serious ways. On the other hand, genuinely random samples are very difficult to get. Even with theoretically perfect random samples, surveys have many other sources of errors in real world. We live in a world of radical uncertainty, and must often act on basis of very poor information – both in quality and quantity.

In situations of uncertainty, decisions place much more reliance on intuition, rather than rational choice.

8: Causality and Regression Models

This should be a BLURB about chapter 8

8A Flawed Foundations of Econometrics

The central goal of this Lecture 10A on “[Real Statistics: An Islamic Approach](#)”, is to establish that standard methodologies currently in use in statistics/econometrics, particularly regression, are built on the wrong foundations, and incapable of generating knowledge. To understand WHY this is so, it is necessary to study the flawed philosophy of science used to build these foundations.

While the car is functioning well, one does not usually open up the engine. But when the car breaks down, it becomes necessary to open it up to see what is wrong. This is the situation today, as the failure of econometric models manifested itself in the global financial crisis, as well as many other occasions. The tragedy is that these same failed models continue to be used today; no serious alternatives have been developed. The reason for this is that the methodology used to develop these models is inherently flawed, and incapable of producing knowledge. It is necessary to understand the engine – the philosophy of science underlying statistics and econometrics – to see why this is so. A capsule summary outline of why it is necessary to discuss philosophy of science is given below:

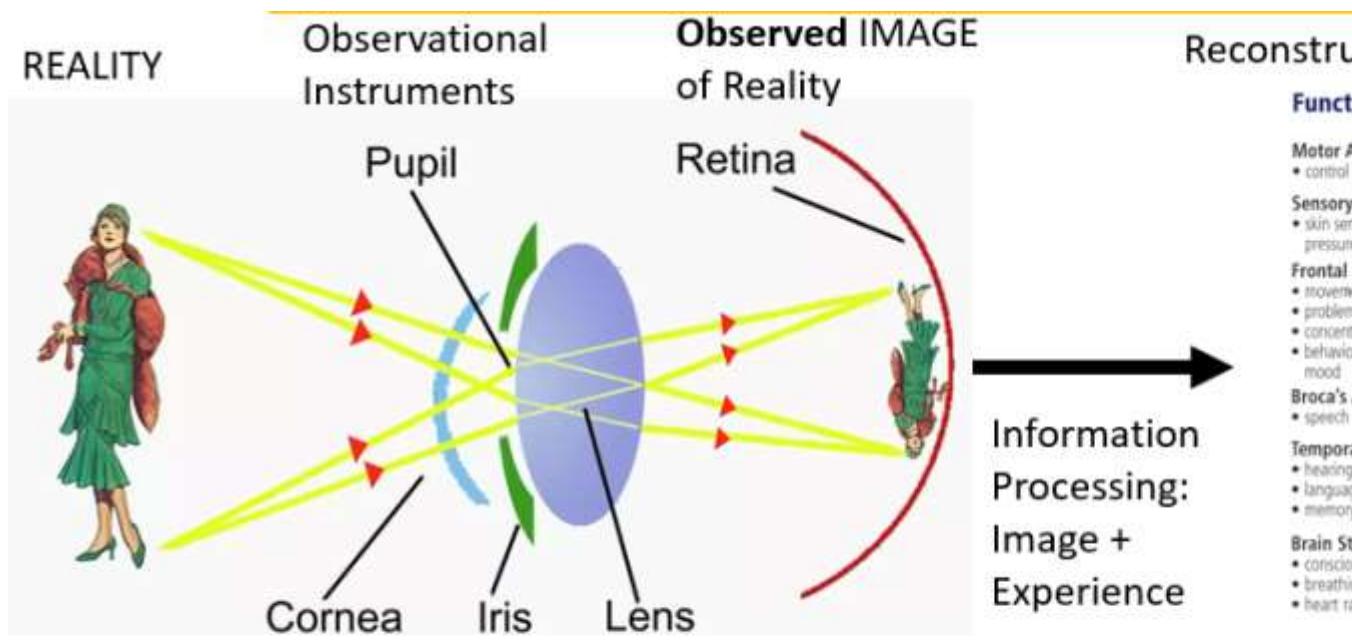
1. Science was imported into Europe from Al-Andalus (Islamic Spain) via the Reconquest in 1492, which made available to the West, millions of books in the libraries of the Islamic Civilization. See “[Is Science Western in Origin?](#)”.
2. These books ended the dark ages of Europe and led to the Enlightenment.
3. Over the next two centuries, there was a tremendous battle between “Science” (Islamic philosophies, science, and other types of knowledge) and “Religion” or Christianity.
4. This battle was won by science, and the “Philosophy of Science” emerged as separate discipline, distinct from science itself. The goal of this philosophy was the prove that science was a source of certain knowledge, and it was the ONLY such source – in particular, all religious knowledge was merely ignorance and superstition.
5. Because of these ideological blinders, the philosophy of science set for itself an impossible task. Therefore, it was not able to make any progress in understanding the true nature of science. To this day, there is massive confusion about what science is, and how it works (for example, see Chalmers “What is this thing called science?”).

6. Mistaken “positivist” understandings of science were used to build the foundations of economics, statistics, and econometrics. Today, it is an urgent need to recognize these flawed foundations, and rebuild these disciplines (and all of the social sciences) on new foundations.

The First Scientist: Ibnul Haytham

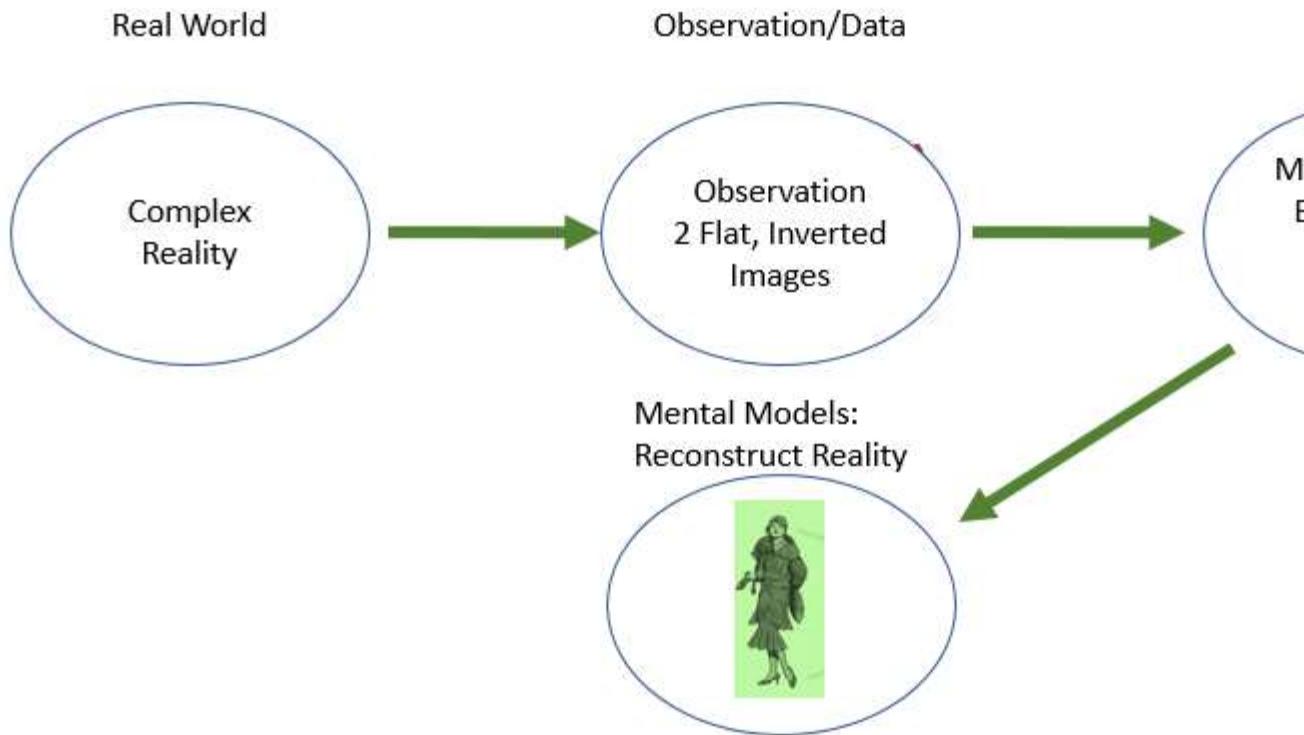
Mathematics, especially the geometry of Euclid, was the first discipline of knowledge established by Greek Philosophers. This was based on taking intuitive certainties as axioms, and then deducing more complex truths by using logical deductions. This is called the axiomatic-deductive methodology. When the Greeks turned to the natural sciences, they attempted to use the same methodology. Unfortunately, this methodology does not work well in this case. For centuries, philosophers were divided on the issue of whether light emanates from eyes to strike the object, or whether light comes from the object to the eye. There were axiomatic-deductive demonstrations for both positions. Ibnul Haytham was the first to use empirical methods to resolve this controversy, laying the basis for the scientific method. It is worth discussing his contribution in detail, because the concept of a “MODEL” emerges from his study. This concept is central to understanding the problems with current foundations of the social sciences. See “[Models & Reality](#)” for further discussion on this point.

The diagram below describes the understanding of vision which Ibnul Haytham came to, as a result of his scientific methods of investigation:

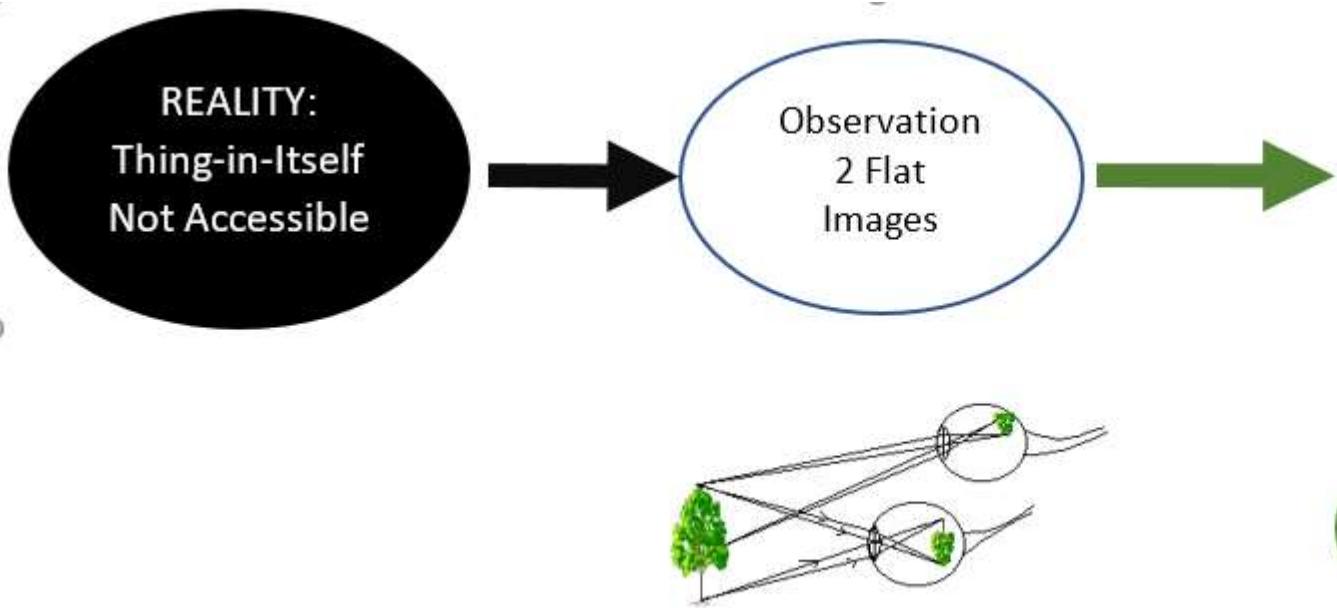


The Process of Understanding

Light from the object (woman) travels in straight lines, is focused onto our retina within our eyes. An inverted image of the woman is formed in the two retinas. Our MIND analyzes the two images and RECONSTRUCTS the external object. What we see directly are the images on the retina, and the picture of reality is created by the MIND based on calculations and past experience in interpretation of such images. Thus we see with our minds, not with our eyes. A schematic sketch of how we see is given in the diagram below:



It is crucial to understand that we do NOT directly see the external world. Our mind recreates a picture of external reality based on clues furnished by the images on our retina, which actually give us an inverted picture of reality. An amazing experiment was performed to show how we see with our minds. A student was fitted with inverting glasses, which make the world appear upside down, and told to keep them on constantly. After a few days of dizziness and disorientation, he learned to see through these glasses without difficulty. THEN the world appeared upside down when the glasses were removed. The mind was able to re-interpret the inverted image and fix it, to enable the person to see the world as it is.



The quest of traditional philosophers was to establish that our mental models of reality matched (or did not match) external reality. Kant argued that this problem was impossible to solve, since we had no access to reality other than by our observations. So, he proposed to CHANGE the problem. Instead of asking whether our mental models matched external reality, he said that we should assess how our mental models are constructed from the observations. After Kant, instead of matching mental models to reality, the focus shifted to matching mental models to observations.

A realist philosophy of science asks us to build models which are closely matched to hidden structures of reality which generate the observations that we can see. However, a nominalist (empiricist, positivist) philosophy is concerned only with building models which provide a good match to the observations, without any concern about reality. The shift from realism to nominalism – for reasons far more complex than Kant's philosophy discussed above – had disastrous consequences, especially for the social sciences. Modern social sciences were created in the early 20th century, based on conscious adoption and imitation of methods of the physical sciences. However, these methods were misunderstood; it was assumed that “science” only deals with observables, and not with unobservables. This has led to deeply flawed foundations for social science. We briefly explain the implications for economics and econometrics.

Impact of Positivist Methodology on Economics & Econometrics

The most famous and widely read methodological essay in economics is Friedman's “The Methodology of Positive Economics”. In this essay, Friedman argues that good models have “bad” (false) assumptions – in fact, “The more wildly inaccurate the assumptions, the better the model”. The meaning here is that if a drastic over-simplification of a complex reality gives good results in terms of providing a good fit to observations, this is the sign of a good model. However, this methodological principle gives us a license to make any assumption we like, as

long is it produces a good fit to the data. This is what results in terrible models in economics, statistics, and econometrics, as we briefly illustrate.

Top ranked economist Lucas writes that “Unlike anthropologists, however, economists simply invent the primitive societies we study.” The “invented society” is populated by “homo economicus” a robot with behavior predictable by mathematical laws. In principle, economists are supposed to check if the results from their artificial models match observed reality. In practice, they rarely bother to do so. See my paper on “Models and Reality: How Models divorced from Reality became epistemologically acceptable”, for details.

In statistics, we start with data on a variable X, observed across time to get observations $X(1), X(2), \dots, X(T)$. We assume, without any justification, that all of these observations are random samples from a common infinite population. If the data appears to “fit” our assumption, this by itself justifies the assumption, without any need of checking the assumption against external reality. We have argued in this course that this leads to defective inference, and we should approach data analysis without making such unjustifiable assumptions, in accordance with a realist philosophy of science.

In econometrics, we go several steps further. Given MULTIPLE data series X, Y, Z, we choose the variable we want to explain; say Y. Then we IMAGINE causes of this variable (say X, Z) and call them the explanatory variables. Next, we IMAGINE that there is a LINEAR relationship: $Y = aX + bZ + \text{error}$. Next, we make ASSUMPTIONS about the errors, and causal relationships between Y, X, Z, error. In particular, we assume that X, Z, are causes of Y and are independent of the error (no causal relationships in either direction). After making all of these unjustifiable assumptions, we do calculations on this basis. If our regression model fits well to the observed data, this is taken as sufficient justification for all of our assumptions. We will show in remaining lectures that this methodology leads to disastrously bad models, which yield hopelessly poor policy implications. Below are some quotes in support of this assertion:

1. JM Keynes: “professional economists … were apparently unmoved by the lack of correspondence between the results of their theory and the facts of observation”
2. Solow: To discuss economic theory seriously with Lucas & Sargent is like discussing cavalry tactics at Austerlitz with a madman who believes himself to be Napoleon Bonaparte. Instead, I prefer to just laugh!
3. Romer: modern macro theories give wildly incorrect predictions and are based on fundamentally flawed doctrines, beyond the possibility of repair.

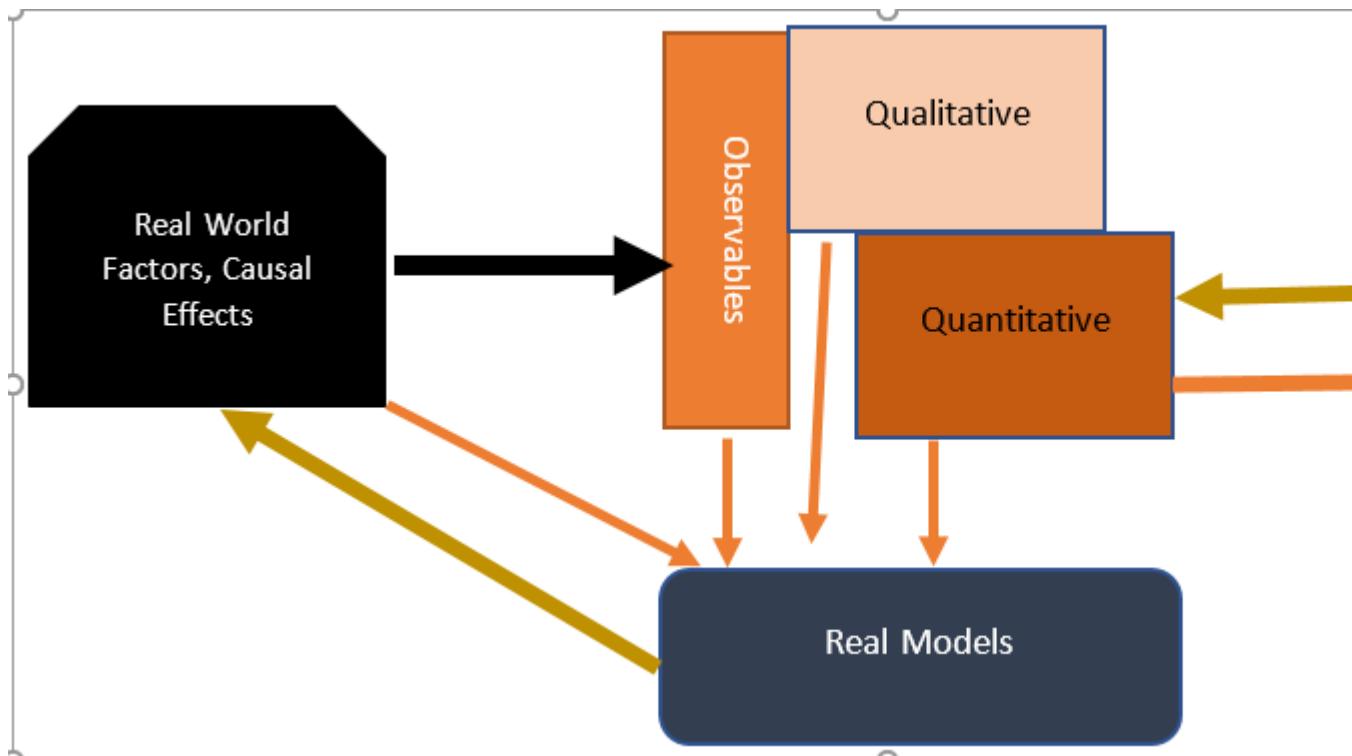
LINKS to related materials: Flawed foundations of social sciences: [The Emergence of Logical Positivism](#) (shortlink: <http://bit.do/azelp>) Sources for above quotes, and additional [Quotes Critical of Economics](#) – <http://bit.do/azquo> More details about above arguments [How Economic Models became Substitutes for Reality](#). { <https://ssrn.com/abstract=3591782> }. Word Document with this writeup: [rsia10A Phil Science.docx](#)

8B Real Models Versus Econometric Models

This is the second in a sequence of lectures about regression models and methodology. These regression models are central in econometrics. Current methodology underlying these models makes them completely useless for learning about the real world. To learn how and why, we first discuss the differences between nominalist and realist methodology for science.

Underlying Philosophy of Science

Many important structures of the real world are hidden from view. However, as briefly sketched in L10A, current views say that science is only based on observables. Causation is central to statistics and econometrics, but it is not observable. As a result, there is no notation available to describe the relationship of causation between two variables. We will use $X \Rightarrow Y$ as a notation for X causes Y. Roughly speaking, this means that if values of X were to change, then Y would have a tendency to change as a result. This is not observable for two separate reasons. ONE because it is based on a counterfactual. In another world, where the value of X was different from what was actually observed in our current world, this change would exert pressure on Y to change. TWO X exerts an influence on Y, but there are other causal factors which are also involved. Thus Y might not actually change in the expected direction because the effect of X might be offset by other causal factors which we have not accounted for. For both of these reasons, causality is not directly observable.



The diagram illustrates the dramatic difference between current conventional econometric models, and the proposed replacement by real models we will introduce in this

course. The real world consists of “factor” or “variables” and causal effects which relate these to each other. Some of these variables within the real world are observable. These observables are divided into the qualitative (like colors, and countries) and the quantitative (measurable numerically). Econometrics attempt to create models which provide the best possible fit to the quantitative observables. Real models are very different, both in terms of how they are constructed, and what they aim to achieve. Because the foundations of econometrics are thoroughly positivist, we lack the language, notations, and symbols, required to talk about real world factors and causal effects. It is the extreme ambiguity introduced by this lack of conceptual clarity which allows econometrics to function: we write the regression model in one way, and interpret it in another, without noting the difference. Accordingly, before explaining real models, we pause to introduce some terminology.

Achieving Conceptual Clarity

The standard approach to statistics and econometrics is based on a huge number of confusions. The same word is used for many different concepts. To clear up these confusions, we need to develop new language and notations. We start by distinguishing between three different types of ideas:

1. O-concepts refer to the Observables.
2. M-concepts refer to a model for the data
3. R-concepts refer to the Real World

R-concepts refer to factors in the real world, and causal effects which link them. For example, Household income could be one of the factors which causally influences consumption decisions. Let us use HI^* and HC^* to denote real world household income and consumption for some particular household. We will use \Rightarrow to denote causation: $HI^* \Rightarrow HC^*$. A household income-expenditure survey obtains measures HI and HC of HI^* and HC^* . We will use this notational convention to distinguish between real world concepts and their observable counterparts. An econometric model attempts to find a model which fits data on HI and HC. A real model uses the data on HI and HC as clues to tease out causal relationships within real world variables HI^* and HC^* .

To illustrate how real-world models are constructed, we go through a hypothetical example close to reality. We start with a hypothesis about a real world causal relationship; for example, $HI^* \Rightarrow HC^*$. Causal relationships are unobservable, so no direct confirmation is possible. However, examination of data on HI and HC can provide indirect evidence confirming or disconfirming the hypothesis. There are three main possibilities: $HI^* \Rightarrow HC^*$, $HC^* \Rightarrow HI^*$, and $HC^* \wedge HI^*$. The last is the standard symbol for independence, but since this is not generally available, we will also use || double vertical bars as a replacement notation: $HC^* \parallel HI^*$ means that the two variables are independent – neither causes the other. Note that there would be many other possibilities, such as bidirectional causality, or causal effects mediated through intervening variables, but we are considering the simplest possible cases, to start with.

The data can provide us with evidence regarding these causal relationships. If we see large variations in HI* and very little in HC*, we would be tempted to reject the causal hypothesis that $HI^* \Rightarrow HC^*$. This might be the case in an ideal Islamic society, where everyone follows simple lifestyles, regardless of income levels. If, on the other hand, we see that consumption levels increase with income, this would suggest that our hypothesis may be true. But, we always need to check for reverse causation. Suppose for example that people are accustomed to different lifestyles, and they earn to support their lifestyle. Those who desire higher consumption levels will be driven to earn higher incomes. In this case the causal direction will be the reverse: $HC^* \Rightarrow HI^*$. There are many different ways that we can judge the direction of causation, according to availability of data, or using experiments which vary income.

The central point we are trying to make here concerns the difference between real models and econometric models. Econometrics models are confined to the OBSERVED data HC and HI. Real models ALWAYS go beyond the observed data, and involve causal hypotheses linking the unobservables HC* and HI*. Real models can never be proven or disproven, but data can provide support or disconfirming evidence. In this regard, the data is suggestive, never conclusive, for a number of reasons. First, what is observed is an imperfect measure of the underlying real variable. Second, causal effects may be suppressed in the sample due to operation of other factors about which we have no knowledge. For example, we might observe a sample where consumption is identical but income levels vary greatly, and conclude that the causal hypothesis $HI^* \Rightarrow HC^*$ is not valid. However, we may find that data is for a population of migrant workers, who send all their savings back home to their families, while minimizing personal consumption levels to what is barely necessary. Here, another factor is operating to suppress the causal effect which would appear in its absence.

Regression: Most widely used model

Fisher introduced the idea of making an assumption that data is a random sample from a hypothetical imaginary distribution, in order to simplify data analysis. Regression extends this idea to two or more variables. It is based on LARGE number of FALSE assumptions. As we have seen, a nominalist methodology has no difficulties with false assumptions. It does not matter if Model doesn't correspond to Reality. We only check FIT between model & data. In contrast to this, REALISM says that False assumptions lead to false results; that is our premise in this course on Real Statistics.

A Standard Course Introduces Assumptions of the regression model with minimal explanation. The goal is never to analyze or understand these assumptions, or to assess if they are true. The assumptions just provide us with the mathematical tools required to do data analysis. The standard course USES the assumptions to do complex mathematics required to setup and estimate regression models. A whole SEMESTER of work is involved in learning the MECHANICS of regression. Since nearly all regression models are false, all of this is useless. In this course, we will study Regression Models from OUTSIDE. That is, we will discuss how regression models are setup and estimated, without discussing the mechanical and mathematical details.

Causality:

As discussed earlier, causality is not observable. We can see that Y happens after X, and but not that Y happens because of X. Because of this, the positivist methodology of econometrics makes no mention of causality. Yet, causality is central to understanding regression models, and also, why they fail. We have already introduced the notation that “ $X \Rightarrow Y$ ” reads “X causes Y”. What does this mean? Changes in X lead to changes in Y. The relationship is DIRECT. If we have $X \Rightarrow W \Rightarrow Y$ this is NOT equivalent. In this situation, W is a MEDIATOR – it mediates the relation between X and Y. Other causal factors may be present. $X \Rightarrow Y$ and $Z \Rightarrow Y$ is possible and often the case. Other chains of causation may be present. $X \Rightarrow Y$ and ALSO $X \Rightarrow W \Rightarrow Y$ can BOTH hold. However, we will assume that there is no circular causation. We cannot have $X \Rightarrow Y$ and $Y \Rightarrow W \Rightarrow X$. Although such situations may be possible, we ignore them for simplicity in our initial approach to causality. With this notation in place, we can now discuss the assumptions of the regression model.

Assumptions of the Regression Model

Regression starts by identifying a dependent variable Y, which we wish to “explain”. The regressors are a set of variables X_1, \dots, X_k to be used in explaining Y. Since the notion of “explain” is never explained, the meaning of these fundamental assumptions never emerges clearly. In this lecture, we will only consider the case of a single regressor X. The key assumption is one of a causal relationship between X and Y: $X \Rightarrow Y$. This is referred to as the Exogeneity of X, but there is no real understanding of what this means. Next we assume a LINEAR relation between Y and X, and some other factors which are not known:

$$Y = bX + F_1 + F_2 + F_3 + \dots + F_n$$

If we aggregate the unknown factors into an error term, we get the regression model: $Y = bX + ErrY$. An important assumption is that $ErrY$ is INDEPENDENT of X. We will use $=/ >$ to mean “does not cause”. Then the independence assumption $X \parallel F_i$ means $X =/ > F_i$ and also $F_i =/ > X$, where F_i are the unknown factors which go into the error term. To run a regression, we will need multiple data points. Then, an additional assumption is that $ErrY$ is independent across time or sector. We also have the standard Fisherian assumption: $ErrY$ is random sample from common distribution. If we let M be the common mean of this unknown distribution, we can write the regression equation as $Y = M + bX + (ErrY - M)$. In this equation we have a constant term, and the new error term is $ErrY' = ErrY - M$. This error term has mean 0.

1. What is the BASIS for these assumptions? NONE.
2. Can we expect them to be roughly valid? NO!

These are all INCREDIBLE & BIZARRE assumptions – it is almost impossible to think of situations where they would hold. Edward Leamer analyzed the regression model and stated The Axiom of Correct Specification: Regressions produce good results ONLY IF ALL ASSUMPTIONS HOLD. Failure of any one of the assumptions can lead to dramatically wrong conclusions. Many examples of such failure can be demonstrated in highly regarded papers published in top journals. The biggest problem is that regression models create confusion about

causal effects in minds of students. This makes it impossible for them to use data to arrive at sensible conclusions about reality.

Yule fitted the first regression model to data in 1896, and failed to reach any conclusions regarding the problem he attempted. Hundred years later, in a centenary article, Freedman (1996) wrote that we have tried this methodology for a century, and it has failed to produce good results. The time has come to abandon it. This is precisely our point of view. Regression models have failed, and should be abandoned. In this chapter, we will describe the methodology and its failures. Later we will develop alternative, superior methodologies, based on a REALIST approach, which does not allow to make false assumptions freely, to fit the data. With this as background, we now turn to regression models.

Regression: Fitting lines to data

For the two variable case, regression involves fitting a line to data, creating a smooth and simple relationship which approximates a complex and cluttered cluster of points on a graph of the data. In context of REAL STATISTICS – we must ask WHY?

1. Why are we fitting a line to the data?
2. What does the fitted line MEAN?
3. How do we calculate which line fits best?
4. What will be done with the regression analysis?

In real world situations, there are a number of possible DIFFERENT uses for fitting lines. The interpretation of the LINE depends greatly on real world context which generates the numbers, and the PURPOSE for which this analysis is being done.

GENERALLY SPEAKING: Regression methodology is only ONE of many possible ways of fitting lines. VISUAL FITS are USUALLY an excellent & SUPERIOR option. We will now illustrate these abstract concepts by one simple example. More examples will be given in later lectures.

Measuring Serum K levels in Blood:

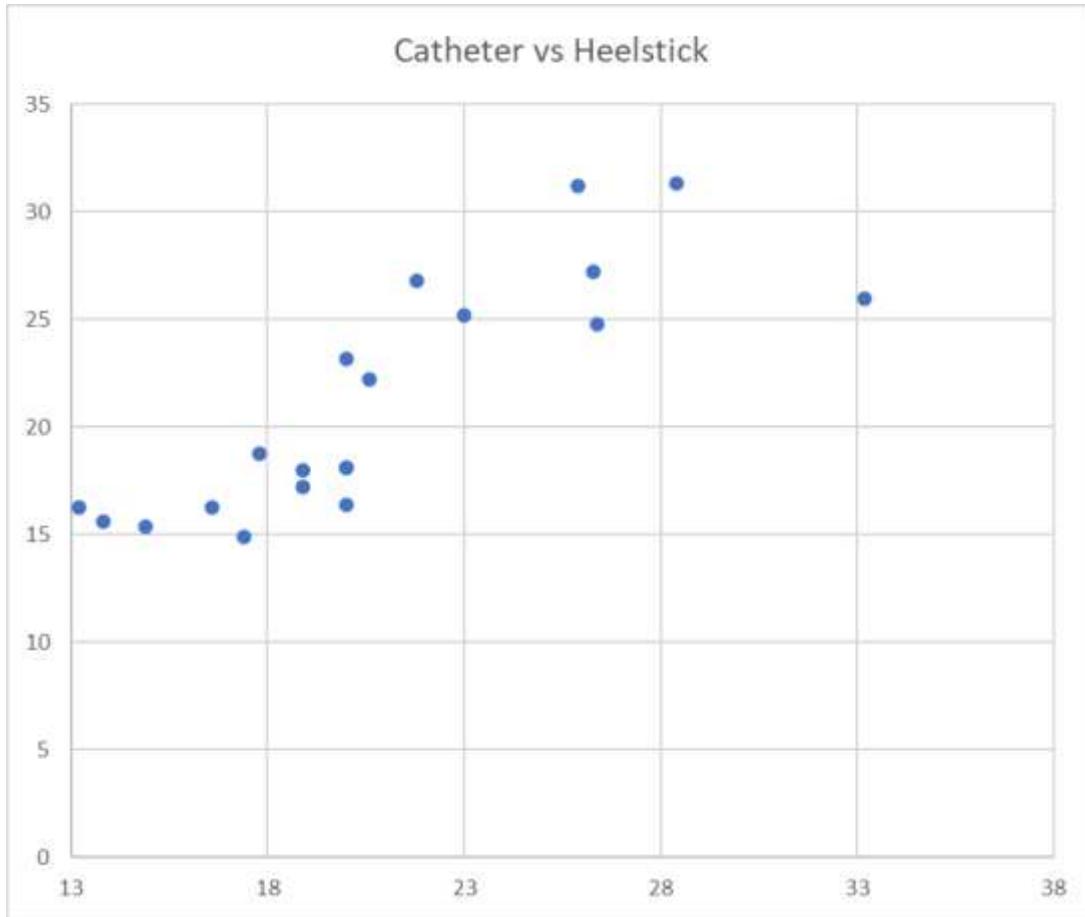
The data we analyze is two ways of measuring Serum kanamycin levels in blood. Samples were drawn simultaneously from an umbilical catheter and a heel venipuncture in 20 babies – this is a real data set taken from Kaggle. The data set and a graph of the data are given below:

| Baby | Heelstick | Catheter |
|------|-----------|----------|
| 1 | 23 | 25.2 |
| 2 | 33.2 | 26 |
| 3 | 16.6 | 16.3 |

Author Last Name/Book Title

| | | |
|----|------|------|
| 4 | 26.3 | 27.2 |
| 5 | 20 | 23.2 |
| 6 | 20 | 18.1 |
| 7 | 20.6 | 22.2 |
| 8 | 18.9 | 17.2 |
| 9 | 17.8 | 18.8 |
| 10 | 20 | 16.4 |
| 11 | 26.4 | 24.8 |
| 12 | 21.8 | 26.8 |
| 13 | 14.9 | 15.4 |
| 14 | 17.4 | 14.9 |
| 15 | 20 | 18.1 |
| 16 | 13.2 | 16.3 |
| 17 | 28.4 | 31.3 |
| 18 | 25.9 | 31.2 |
| 19 | 18.9 | 18 |
| 20 | 13.8 | 15.6 |

We graph this data in an X-Y scatterplot:

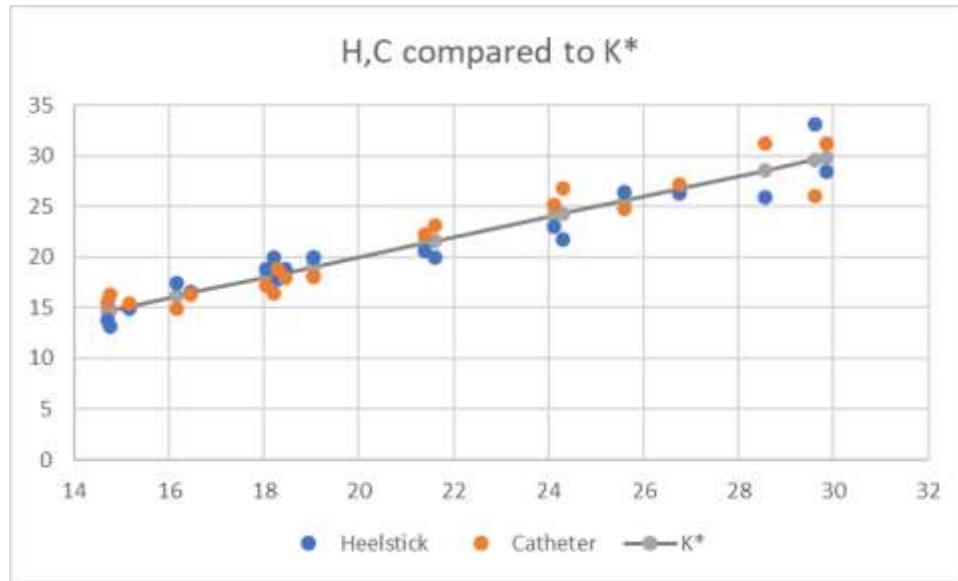


The data and the graph show a rough correspondence between the two measures, which is what we expect to see. Both Cath and Heel are measures of the same unknown quantity K, the serum K levels in the blood of the baby.

The data is relevant to the analysis of a meaningful question. Are the two measures equivalent? Do they provide us with an accurate measure of the true Serum Kanamycin levels (K) in the baby's blood? The unobserved real variable K generates the two measures C and H: $K \Rightarrow C$ and $K \Rightarrow H$. Because both are caused by K, we expect to see high correlation between the two. BUT there is no causal relationship: $C \Rightarrow H$ and also $H \Rightarrow C$. Neither measurement causes the other. A reasonable representation of the underlying structure is: $C = K + Err_C$ and $H = K + Err_H$. This is A GOOD model because it matches underlying unobserved reality. The errors are meaningful – they represent the errors created by assuming that the sample is representative of baby's blood. As we will soon see, regression analysis cannot use either of these correct structural models because they involve the unknown and unobserved K.

Before turning to a regression, we do a little common-sense analysis of the data. It is well known that given two erratic but independent measurements, an average of the two will give us a more stable and accurate measure. While K cannot be measured, if we define K^* as the average of the two measures we have – $K^* = (C+H)/2$ – then K^* will provide a better approximation to the underlying true K than either of the two measures C and H. Using this idea as the basis, we

construct a table of values for K^* and also plot the errors – the deviations from K^* - of the two measures below.



The X-axis is our estimated value K^* of K . On the Y axis we plot K^* and compare it with the two measures H and C . As we can see, both measures are closely matched to K^* , so the data conforms to our intuitions regarding the two measures H and C , and the underlying unobserved K . There are few things we can learn from this data analysis. First we present the “errors” in the measures C and Y and measures of K^* . Note that these are NOT the TRUE errors, because the true value of K is unknown. Instead, we analyse $\text{Err}C^* = C-K^*$ and $\text{Err}H^* = H-K^*$

| K^* | ErrH | ErrC |
|-------|------|------|
| 14.7 | 0.9 | -0.9 |
| 14.75 | 1.55 | 1.55 |
| 15.15 | 0.25 | 0.25 |
| 16.15 | 1.25 | |
| 16.45 | 0.15 | |
| 18.05 | 0.85 | |
| 18.2 | -1.8 | 1.8 |

| | | |
|-------|------|------|
| 18.3 | 0.5 | -0.5 |
| 18.45 | 0.45 | 0.45 |
| 19.05 | 0.95 | 0.95 |
| 19.05 | 0.95 | 0.95 |
| 21.4 | 0.8 | -0.8 |
| 21.6 | 1.6 | -1.6 |
| 24.1 | 1.1 | -1.1 |
| 24.3 | 2.5 | -2.5 |
| 25.6 | -0.8 | 0.8 |
| 26.75 | 0.45 | 0.45 |
| 28.55 | 2.65 | 2.65 |
| 29.6 | -3.6 | 3.6 |
| 29.85 | 1.45 | 1.45 |

The three errors highlighted in red are the largest errors. All other errors are within 2 units of K*. From this analysis, we could conclude that the two measures are aligned, and both come within 2 units of the true K about 85% of the time. Occasionally, larger errors like 2.5 or even 4 can occur. One could go further by examining the particular cases of large errors to try to identify the source of the error. A real analysis always goes beyond the data, to try to understand the real world factors which generate the observations. Next, we turn to a regression analysis of this data set.

External Regression Analysis

A conventional course in regression analysis does the following:

1. Learn assumptions of regression
2. Use to create mathematical & statistical analysis
3. Learn how to estimate regression models, & properties of estimators & test statistics
4. Use this theory to interpret results of regression

Our point of view here is that the assumptions are nearly always false. As a result, the results are nearly always useless. So there is no point in learning all of this machinery, which take a lot of time and effort. Instead, we will teach regression from an external perspective. Running a regression involves doing the following tasks:

1. Choose Dependent Variable Y
2. Explanatory variables X
3. Goal explain Y using X.
4. Feed data to computer – we don't investigate what happens inside the computer.
5. Get regression output
6. LEARN how to INTERPRET output.

A huge amount of time can be saved by learning how to drive the car, without learning the details of how the engine works. Accordingly, we proceed by running a regression analysis of the data on C and H.

We immediately run into a problem. Which of the two should be a dependent variable, and which should be the explanatory variable? Actually, the causal structure shows that both are dependent, while the independent variable is K. However, regression analysis does not allow the use of unobservables – we have no data on K. The nominalist philosophy says that we should use variables which are unobservable. Accordingly, there are only two possibilities, we can run a regression of H on C or of C on H. Both are wrong because both embody the wrong causal hypotheses, and end up giving us wrong and misleading results. We will examine the regression of H on C in greater detail. There are vast numbers of programs which we can use to run regressions; they all give similar results. Below, I present the results obtained from EXCEL:

| SUMMARY OUTPUT | | | | | | |
|------------------------------|-----------------------|--|--|--|--|--|
| B on C | Heelstick on Catheter | | | | | |
| <i>Regression Statistics</i> | | | | | | |
| Multipl e R | 0.8324 | | | | | |
| R Square | 53 | | | | | |
| Adjusted R Square | 0.6929 | | | | | |
| 21 | 0.6759 | | | | | |

| Standar d Error | 2.9042 64 | | | | | | |
|-----------------|---------------|-----------------|--------------|--------------|-----------------|--------------|-----------|
| Observ ations | 20 | | | | | | |
| ANOVA | | | | | | | |
| | Df | SS | MS | F | Signifi cance F | | |
| Regression | 1 | 342. 684 | 342. 684 | 2765 | 40.6 | 5.29E- 06 | |
| Residua l | 18 | 151. 8255 | 8.43 4749 | | | | |
| Total | 19 | 494. 5095 | | | | | |
| | Coeffi cients | Stan dard Error | t Stat | P- value | 95% | Lower | Upper 95% |
| Intercept | 4.2101 12 | 2.69 0918 | 1.56 4563 | 0.13 5096 | - 1.4433 | 3522 | 9.86 |
| Cathete r | 0.7869 92 | 0.12 3469 | 6.37 3982 | 5.29 E-06 | 0.5275 93 | 1.04 6392 | |

This output is interpreted as follows.

Correlation & R-squared: Multiple R = 0.832453 ,R Square = 0.692978

In the nominalist theory, the most important aspect of regression is how well the model fits the data (not how well the model approximates truth). For this purpose, the correlation and its square are the most important measures in use. These measures the association between the dependent variable Y and the regressors, under stringent assumptions about the data. This measure can give HIGHLY misleading results if assumptions are violated. Two variables which have no relation to each can have very high correlation. Also, two variables which are strongly related can have zero correlation. In this particular case, the two measures H and C are linearly related, so the correlation is a suitable measure of their association. The 83% correlation is high, and shows a fairly strong relationship between H and C. The R-square is SQUARE of correlation: $69.2\% = 83.2\% \times 83.2\%$. This has the following standard Interpretation: 69.2% of the VARIATION in H variable can be explained by the C variable. This concept is meaningful ONLY under very strong assumptions rarely satisfied in real world applications. This comes

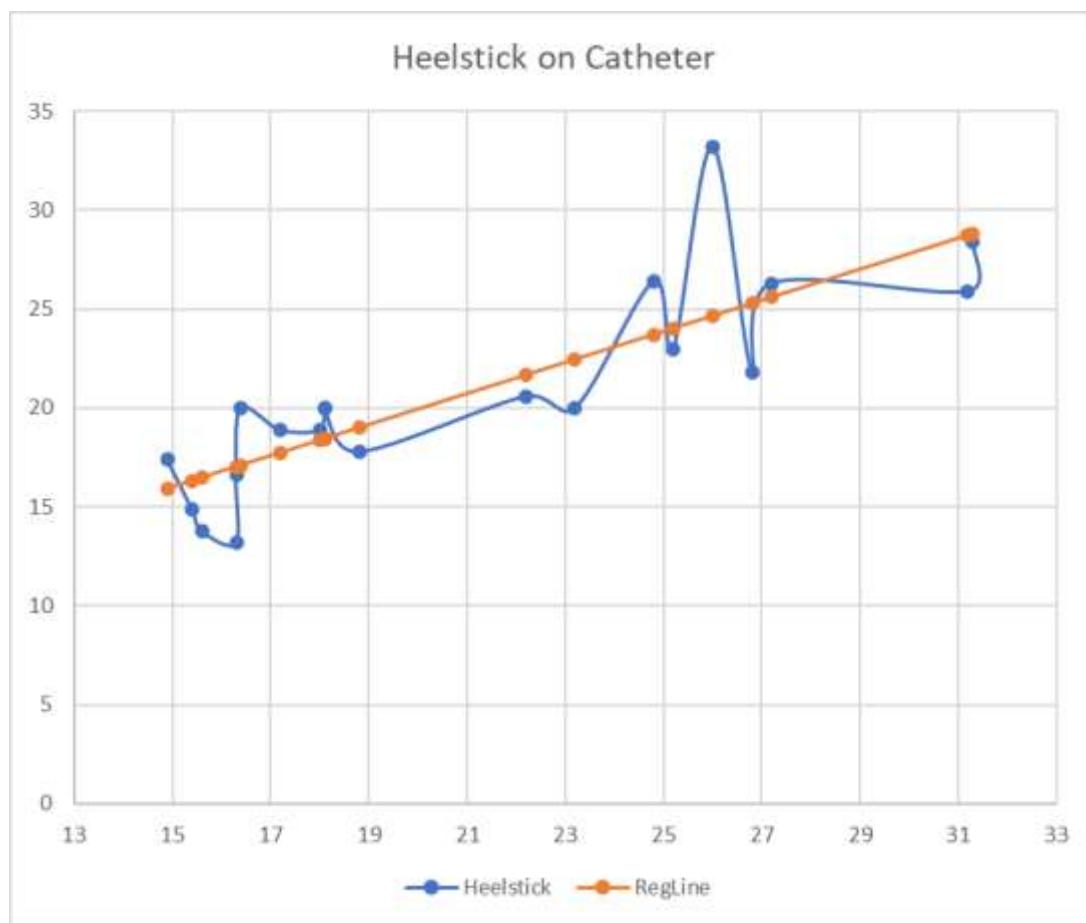
from ANOVA analysis due to Fisher pursuing racist goals. Fisher wanted to explain variations in children use parents genes as causal factors, and to separate the contribution of genetics from the environment. However, this methodology does not actually succeed in achieving this goal.

Central Object of Regression is to FIT A LINE:

The most important goal of regression is to approximate the dependent variable as a linear function of the independent variables. The regression output estimates that H is a linear function of the C:

$$H = 4.21 + 0.79 C + \text{err}$$

A graph of this relationship is given below. The orange is the regression line, while the blue is the actual data.



Visually, we can see that the data fits reasonably well to the line. This reflects the high correlation of 83%. However, this apparent linear pattern in data is not a good reflection of the TRUE relationship between these two variables H and C. The regression estimates don't make sense. They have been derived under assumption that C is fixed, and H is caused by C. But this is not true.

A standard interpretation of this regression relationship would be that if we change C by one unit, then H would change by 0.79 units. This is NOT valid, because C is dependent on K and an Error ErrC. If we change ErrC – that is, the error of measurement – this will cause a change in C but not in K. Since K is the cause of H, changes in the error of measurement of C will not affect H. So the interpretation of 0.79 as the effect of changes of C on H is not correct. Similarly, none of the other statistics generated by regression make any sense, because the assumptions of the regression model are not satisfied.

The F-statistic for the Regression:

F 40.62765 Significance F 5.29E-06

We will end this lesson with an interpretation of the overall F-statistic for the regression as a whole. This is similar to the R-squared in being a measure of the overall goodness of fit between the data and the regression line. First we recall the meaning of the p-value, which is also called the significance level.

Significance Level: Suppose our null hypothesis is that a coin is fair. We flip it 100 times and observe 70 heads. The p-value measures how much this observation is in conformity with the null hypothesis. The p-value is defined to be the probability of the observed event, together with ALL equal or more extreme events. Here more extreme means having lower probability. If X is the number of heads in 100 with 50% success probability on each trial, then the p-value of the observation 70 is $P(X \geq 70) + P(X \leq 30)$ which is $2 * 3.92507E-05$, an extremely low probability. This means that the event is highly unlikely under the null hypothesis. Because there is a high degree of conflict between the observed event and the null hypothesis that the coin is fair. So, we can reject the null hypothesis.

Now, we go back to the interpretation of the F-statistic. This is a test for the null hypothesis that all coefficients in the regression are 0. That is $a=0$ and $b=0$, so that there is no connection between the two measures H and C. The statistic of 40 with significance of $5.3E-06$ means that this is highly unlikely. So, we can reject the null hypothesis of no connection between H and C. While the inference is correct (that is, there is a strong relationship between H and C), the reasoning which leads to this conclusion is wrong. Also, the p-value is meaningless, since the assumptions on which it is based are not valid.

ENDNOTES: Writeup of this lecture in WORD:[RSIA10B Lines.docx](#)

8C: Causality Defined & Compared with Regression

Defining Feature of Real Statistics:

We can define “Real Statistics” as being the search for causal relationships. There exists an enormous amount of confusion about what exactly is a causal relationship. We will take a simple and practical approach, developed by Woodward in his book on “Making Things Happen”. Given a collection of variables $X, Y, Z_1, Z_2, \dots, Z_n$, we will say that X is a cause of Y if we can create changes in Y by changing the values of X. This is a “practical” definition in the sense that if we learn about causal relationship, we can create changes in the world around us.

Why is there confusion about causality?

Children (and animals) are born with the ability to learn about causal relationships, and to use them to bring about desirable changes in their external environment. Since understanding of causation is built deeply into us, ordinary people find it difficult to understand why philosophers are so confused about causality. We need to discuss this issue because dominant methodology of statistics and econometrics is built on foundations of a philosophy (logical positivism) which is a source of enormous confusion. However, students should not worry if they fail to understand the source of this confusion. The point of trying to explain this is to explain why conventional statistics and econometrics is wrong. It does not matter for learning real statistics.

Confusion about causation was created by David Hume, who noted that we can only observe sequences of events – B follows A – but we cannot observe the underlying causal connection that A caused B to happen. This idea of Hume is actually a mistake created by his misunderstanding of the sources of human knowledge. Unfortunately, his mistake became embodied in the heart of the philosophy of “logical positivism” and became widely accepted. In the early 20th Century, all of the social sciences, including statistics and econometrics, were created on the basis of the false foundations of this philosophy. This is the reason why these disciplines have been singularly unsuccessful in leading to increased welfare for mankind as a whole.

At the heart of logical positivism is the idea that reliable knowledge comes only from observations and logic. On the surface, this seems like a straightforward idea. But when “observations” is restricted to observations of the external world (and not our internal experience), this leads to huge mistakes. For a more detailed discussion of these mistakes, see “The Emergence of Logical Positivism”. Causality is defined by manipulation: changing the value of X leads to changes in the value Y. Quite often, this manipulation is not actually possible; in such cases, the definition is based on a “hypothetical” experiment, where we imagine changing the value of X, and seeing how this would impact on Y. By insisting that knowledge is only of what we actually see, and we cannot have any knowledge of what was not observed, this kind of hypothetical experiment is ruled out by logical positivism. This is why the concept of causation cannot be understood by positivists.

Causation: Complexities, Technicalities, and Subtleties

The philosophical confusion about causation is partly because causation is itself complex. Here, we will take a simple and practical approach to the topic, as formulated by Woodward in his book on “Making Things Happen”. The key concept is that X causes Y if we can use changes in X to create changes in Y. It requires some work to make this precise.

Deterministic & Probabilistic Cause: First, we must distinguish between deterministic and probabilistic cause. A deterministic cause is when a change in X actually causes a change in Y. A probabilistic cause is when a change in X leads to a change in the probabilities of occurrence for Y. A simple example of a probabilistic cause is something like COVID

Vaccination. Suppose that the probability of catching COVID is 60% in some reference population. Suppose that vaccination lowers this probability to 10%. This is an example of how vaccination is a probabilistic cause of not getting COVID. In a population of 100 people without vaccination, about 60 will get COVID while 40 will not. Among 100 vaccinated people, 10 will get COVID while 90 will not. Probabilistic causation cannot be detected by looking at individuals. We will find all four types of people: with vaccine and COVID, with vaccine but COVID free, and without vaccine with COVID, and without vaccine and COVID free. It is only the proportions of COVID free people in the two populations which show the probabilistic effectiveness of the vaccine.

Direct Cause: Next, the concept of *direct cause* is of fundamental importance. The idea is that changes in X directly cause changes in Y, without the influence of any intervening variables. We will represent this symbolically as $X \Rightarrow Y$, or $Y \Leftarrow X$. This requires some care to make it precise. Suppose we have a collection of variables under study: $X, Y, Z_1, Z_2, \dots, Z_n$. To assert that X is a direct cause of Y means that for some configuration of values $Z_1=v_1, Z_2=v_2, \dots, Z_n=v_n$, some change in X from $X=\text{observed value}$ to $X=\text{other potential value}$, while cause a corresponding change in Y (or in the probabilities of Y outcomes). The reason we must fix the other variables is to prevent them from interfering with the power of X to change Y. We do not require that all changes in X lead to changes in Y – only that there should be some way to change X so as to create a change in Y. We also do not require X to have the power to influence Y in all environments (as defined by values of Z_1 to Z_n). It may be that for some configurations of values of the Z-variables, X is powerless to change Y. To say that X is a direct cause of Y, we only need some particular configuration of values Z such that changes in X and affect Y.

Indirect Cause: The word “indirect” is ambiguous and can have several meanings. However, we will use it in only one way: X is an indirect cause if it is linked to Y by a sequence of direct causes. For example, $X \Rightarrow Z_1, Z_1 \Rightarrow Z_2$, and $Z_2 \Rightarrow Y$. In the late 20th Century, Judea Pearl created a new approach to causality, which broke out of the mindset created by positivism. In his terminology, a direct cause ($X \Rightarrow Y$) is a (causal) parent of Y. An indirect cause ($X \Rightarrow Z \Rightarrow Y$) is an “ancestor” of Y. Note how fixing variables allows us to differentiate between direct and indirect causes. If $X \Rightarrow Z \Rightarrow Y$ and we fix the value of Z at v, then regardless of how we change X, we cannot create changes in Y. This is because Z is the only channel through which X affects Y, and if Z cannot change then X is powerless to influence Y. In such situations we say that Z screens the effect of X on Z. After fixed Z, X and Y become independent.

The Collection of Relevant Variables: It is important to note that these definitions depend on the specified collection of variables. If $X \Rightarrow W \Rightarrow Y$, but W is not in the set of variables under consideration, then X will appear to be a direct cause of Y; fixing all other variables will not prevent changes in X from affecting Y. Only fixing the value of W will prevent X from affecting Y, but W is not among the set of variables under consideration. This relativity of direct causation to the variable set under consideration cannot be avoided.

The GOAL of Real Statistics: With these technicalities defined, we can sharpen definition of “Real Statistics”. Given a collection of variables X_1, X_2, \dots, X_n , real statistics aims to find the direct causal relationships between any pair of variables in this set. Given all the

direct causal links, we can also find all the indirect causal links because these are just sequences of direct causal links. There are a few important points which follow from this goal.

1. Each causal link represents a real world mechanism. Changing X causes something to happen which leads to changes in Y. This is NOT a relationship between numbers (the observed values of X and Y) but a real-world relationship is which being captured by the numbers.
2. Because we are after discovery of real world mechanisms, we will ALWAYS need to go beyond the numbers to assess causal linkages. Numbers can only reveal correlations, and never causation. Correlations can often be helpful in discovery of causations, but actual manipulation and control of Y using X lead to far more certain discovery of causes. Again, this requires actual interference and action, not just passive observation, of the real world.
3. Given a collection of real variables, the number of causal hypotheses that can link them is so enormous that it is impossible to go in with an open mind and hope to discover something. Instead, we go in with some tentative causal hypotheses based on our knowledge of the real world structures generating the data. Then the data may support, or discredit, our original guess at the causal structure. If the data conflict with our original hypothesis, they will also often give us a clue about a better alternative. These alternatives may often involve expanding the data set beyond the original set of variables under consideration.

A Real-World Example:

We will illustrate all of these abstract concepts in the context of a real world data. This is [a real data set](#) which lists prices (per square foot) of houses (HPsqf), and also the number of convenience stores (#Shops) within walking distance. There are 415 houses, categorized according to the number of stores, which varies from 0 to 10. The following graph provides a picture of the data set:



Each blue dot represents a house. The number of convenience stores (#Shops) is listed on the X-axis and varies from 0 to 10. There is one outlier, an extremely expensive house which has only one store in the neighborhood. In general, housing price shows an increasing tendency with #Shops. That is depicted by the orange regression line which has a positive slope. Running a regression of Housing Prices on #Shops leads to the following output:

This output can be summarized in the following regression equation:

$$\text{HPsqf} = 27.18 + 2.63 \# \text{Shops} + \text{Error}$$

What the regression tells us is that if we increase the number of convenience store by 1, the prices of houses will increase by about \$2.63 per square foot on the average, in the neighborhood of the convenience store. At least, this is what a CAUSAL interpretation of the regression model would tell us. This is what students are trained to believe, even though the regression does not actually provide us with any causal information directly. To understand this, let us first look directly at a data summary, without regression. One convenient graphical representation is given below:

X=#Stores, Bars=Price, Line=#H



Each bar corresponds to the number of convenience stores in the neighborhood, which varies from 0 to 10. The bar is the median price (per sq. ft.) of houses having that number of convenience stores within walking distance. The orange line is the total number of houses within this category. The graph shows a generally increasing trend in price per sq. ft. corresponding to increasing numbers of shops. The orange line shows that the largest number of houses (around 68) have 0, and a similar number have 5 stores within walking distance. From 5 onwards the number of houses declines. There are only around 10 houses in the category 10 – which means 10 convenience stores within walking distance.

The first thing to understand is that this is the data – this is all the information provided to us by the data. There is no more. That is, the regression analysis does not magically create more information for us. In fact, all the of the “additional” information provided by regression is created by the assumptions of the regression model which are added to the data. In general, and in this particular, these assumptions are almost surely false. So, regressions create an illusion of precision, which is not actually part of the empirical evidence available to us.

Next, we consider the evidence provided by this data. It does seem to be the case that housing prices tend to increase with the number of shops. It is also of some interest that the number of houses within each category is decreasing, at least after 5 shops. The smallest number of houses have 10 shops within walking distance. The question is: is this a causal relationship? Is it true that if the number of shops increase, then housing prices per square foot would go up?

Regression methodology teaches students to believe so. Coming out of a standard econometrics course, students would look at the above regression and conclude that if one additional shop opens up, the prices of houses in the neighborhood would tend to increase by about \$2.63 per sq. ft. on the average. This is completely false and misleading. The data do not provide us with any such evidence.

From the real statistics perspective, the data informs us of a correlation between #Shops and HPsqf. This is a clue to a possible causal relationship between the two variables. There are three possible causal sequences which could lead to such a correlation: #Shops \Rightarrow HPsqf, HPsqf \Rightarrow #Shops, or Z \Rightarrow #Shops and Z \Rightarrow HPsqf, where Z is some unknown common cause which affects both of the variables we see. How can we find out which of these possibilities (ignoring more complex ones) holds? To learn about causality, we must formulate hypotheses about structures of reality which lead to the observed correlation. Three such hypotheses are formulated below:

1. People look for houses with more convenient shopping.
2. Richer parts of town attract more stores.
3. There is some other factor which attracts stores, and also causes higher house prices.

To give an example of the third hypothesis, suppose that the town is centered around a lake. Lakefront properties are generally more expensive. Also, tourists coming to town generally come to see the lake. As a result, there are a lot of shops on the lakeside (which serve tourists). Then the correlation between housing prices and number of shops is accidental, due to a common cause (lake). How can we differentiate between the three hypotheses? No amount of sophisticated data analysis of the data will reveal any information about this matter. Instead, we must expend shoe leather. We could go and ask questions from three classes of people:

Real Estate Agents: What do people look for when they are shopping for houses? How much importance do they give to the number of shops in the neighborhood? Why are the more expensive houses so highly priced, relative to the others?

Home Buyers: What were the characteristics you were looking for when you purchased the house? How much importance do you place on nearby convenience stores, in terms of purchase decisions? How much more would you have been willing to pay for this house, if there were a few more shops located nearby?

Shop Keepers: What influenced you to open up a shop in this location? Were you looking for proximity to other shops, or proximity to a rich neighborhood?

Acquiring information of this sort is necessary to learn about the causal relationships which create the observed correlation. It is obvious that the data cannot answer any of the questions above, which will provide us with this causal information. This demonstrates how a real analysis, which searches for causes, nearly always goes beyond the data, to the real-world factors which generate the data.

Conclusions:

Data analysis for the purpose of discovering causal relationship differs dramatically from the regression analysis methods currently being taught the world over. Some of these differences are summarized below:

1. Think intuitively, about the real world. This generates initial hypotheses about causal structures, to be tested and verified with data. Often special purpose data will have to be gathered for this purpose.
2. Think about direct versus indirect effects: It is essential to identify the direct effects, since these are the building blocks for the indirect effects. Every hypothesized direct effect corresponds to a real world mechanism (not just a pattern in the data). Given a real world mechanism in operation, there are often many different ways – not just data analysis – to assess its presence and strength.
3. Think causal explanation: Whenever we see a strong correlation between variables X and Y, we can look for an explanation. There are three main causal sequences which explain such correlations: $X \Rightarrow Y$, $Y \Rightarrow X$, or $Z \Rightarrow X \& Z \Rightarrow Y$. Thinking about which of the possibilities holds is not an exercise in data manipulation. This is an exercise in thinking about mechanisms which operate in the real world, relating the variables under study.
4. Think about OTHER relevant factors. Whenever we see a correlation, it may be due to a common cause. We have to apply our real-world knowledge to discover what a common cause may be. Then we may be able to gather data on the common cause, and discover whether our hypothesis of common cause holds. If conditioning on the common cause makes X and Y independent, then the hypothesis is confirmed.

This should show how a real data analysis always involves thinking about real world mechanisms, and not about how to manipulate the data, or to make fancy statistical assumptions about the error terms.

DATA on Prices of Houses and Shops taken
from: <https://www.kaggle.com/quanbruce/real-estate-price-prediction>

8D: Autonomy & Invariance: Causally Defined

As discussed by Hoover in “Lost Causes”, the concept of causality was dropped from econometrics for many decades, although it is currently making a resurgence. Nonetheless the revolution launched by Judea Pearl in the 1990’s has not percolated to econometrics textbooks. As a result, satisfactory definitions of key concepts required to make sense of regression models are not available. In this lecture, we will clarify many confusions about regression models using the newly developed tools of causality.

Brief History of Econometrics: Launched in early 20th Century by Ragnar Frisch, econometric methodology was strongly shaped by the Cowles Commission (CC) in the 1960’s. The CC approach relied on structural equations, which embodied causal information known in advance to the researcher. The goal was estimation of causal effects, and not discovery or assessment of the hypothesized causal structures. The oil shock of the 1970’s led to dramatic

failures of macroeconomic regression models, leading to general distrust of econometric methodology. Two major critiques emerged. The Sims critique argued that hypothesized causal structures were false, and should be abandoned. We should go back to a purely descriptive analysis, looking for patterns in the numbers, without any reference to the real world phenomena represented by these numbers. Directly opposite was the Lucas critique which said that regression models were based on surface relationships between the numbers and ignored the deeper causal structures which drive these relationships. Regression models would fail when economic regimes underwent structural change – precisely when they were most needed. Neither approach has led to successful macro models. Models based on both approaches, as well as more conventional macro models, failed dramatically in the Global financial crisis. The fundamental problem lies in the failure of econometrics to incorporate causal inference correctly.

Autonomy & Invariance: These are two concepts which are of central importance, but cannot be understood without using causality. Autonomy of a regression equation means that the relationship continues to function even if other relationships in the economic system change. For example, a consumption function arises out of stability in the use of income for household purchases. We would expect this relationship to persist over many different kinds of economic change, though not all of them. Ragnar Frisch classified regression relationships into three types: (1) structural (=autonomous), (2) confluent, and (3) spurious. There is substantial confusion in econometrics literature over the meaning of these terms. However, they can easily be understood within the causal framework discussed in the previous lesson.

Causal Explanation: A regression relationship $Y=a+bX$ is **autonomous** if it represents a causal relationship $X \Rightarrow Y$. This causality means that there is a real-world mechanism which operates to transmit the effect of changes in X to the variable Y . It is this mechanism which guarantees the autonomy of the equation. If the world changes in ways which do not affect this mechanism, it will continue to operate, and produce the desired relationship between X and Y . A **confluent** relation is one where $X \Rightarrow W \Rightarrow Y$. Because W mediates the causal effect, shocks to W can disturb the relationship between X and Y . The relationship is genuine, but it is not invariant to systematic changes in W . A spurious relationship occurs due to a common cause. Suppose $Z \Rightarrow X$ and $Z \Rightarrow Y$. The variables X and Y appear related because changes in Z are transmitted to both. If we fix Z , the two would be independent. Despite possibly strong correlation between X and Y , both regressions (X on Y , or Y on X) are spurious.

Invariance: Confusingly, invariance is sometimes used for concepts similar to autonomy. It is better understood in the context of attempting to QUANTIFY a causal effect. Suppose we know that $X \Rightarrow Y$. Next we want to find out the strength of this effect. This is what a regression model does. When we estimate $Y=a+bX$, the estimated coefficient b measures the impact of changes in X on changes in Y . Invariance refers to the constancy of this parameter b . As we will explain below, this search for invariance is an illusion, which side-tracks from the real goals.

Notation: Many authors have suggested that a big source of confusion regarding causality comes from lack of notation to express causal concepts. Accordingly, we will introduce some explicit notation for this purpose. The forward arrow $X \Rightarrow Y$ indicates that X causes Y , as per definition already given in previous lecture. We will use backarrows, as in $Y \Leftarrow A + B X$ to

indicate a quantified causal relationship. Here, not only does X cause Y, but the impact can be quantified by the equation. Nothing prevents multiple causes from affecting Y. Let us say that Z=(Z₁,Z₂,...,Z_n) are other variables which casually affect Y. If they are independent of X, then we can lump the combined influence of all of them into the constant A=A(Z). It is easy to imagine environmental variables E=(E₁,E₂,...,E_k) which could affect the strength of the causal relationship, making it stronger or weaker. The causal factors Z could also play this role. So the constant B is better modeled as a function B=B(Z,E).

Regression: Suppose A(Z)=a₀+a₁Z₁+...+a_mZ_m, where Z₁ to Z_m are observed. We could lump the combined effect of the unobserved variables Z(m+1) ... Z(n) into an error term e and get the standard regression equation: Y = a₀ + a₁ Z₁ + a₂ Z₂ + ... + a_m Z_m + B(Z,E) X + e. If we add the assumption that B remains constant, so the B(Z,E)=B, then we get a standard regression model. In this equation, Z and X play different roles. The main causal relationship of interest is the one between X and Y. The Z's have been put in to take care of the fluctuations of A in the causal relationship between X, so as to get a better measure of the causal effect B. They are sometimes called "co-variates" for this reason. However, there is no way to tell the difference between Zi and X in the regression equation. We need to improve our notation to clearly indicate this difference. We will use square brackets [] to denote a causal factor, and parentheses () to denote a coefficient within a functional form of a causal relationship. With this notation, we have Y <= (A) + (B) [X]. Here the parenthesized (A) and (B) indicate that these are coefficients within a causal relationship, while the square bracket [X] indicates a causal factor. Then the regression equation can be written as Y <= (a₀ + a₁ Z₁ + ... a_mZ_m) + (B) [X], to show the difference between the Z's and the X. The Z's may or may not be causal factors for Y. In this equation, they are added for the purpose of adjusting the constant, so as to allow more accurate estimate of the causal effect (B). It is of importance to note that the idea that B is constant is unlikely to be true. The strength of causal effect is likely to vary with many different factors, as we will see in examples given later.

Mincer's Earning-Education Equation: An equation developed by Jacob Mincer to quantify the effect of education on earnings has acquired central importance in labor economics. This takes the form Earnings (Ern) = a + b Education (Edu) + Other Factors. Let us take for granted the causal effect of Edu => Ern; evidence in favor of this causal hypothesis is overwhelming. Then the earnings equation Ern <= (A) + (B) [Edu] is an attempt to QUANTIFY the strength of the causal effect. To illustrate the difference between covariates and causes, consider the variable Parental Education (PrE). Generally speaking children of educated parents have more education and higher earnings, so PrE can play a role in the earnings equation. Nonetheless, it is clear that it is not a direct cause of Ern. If it affects the constants A, B (which is plausible), then we can write the regression equation as:

$$\text{Ern} <= (\text{A}) + (\text{B}) [\text{Edu}] = (a_0 + a_1 \text{PrE}) + (b_0 + b_1 \text{PrE}) [\text{Edu}] = (a_0 + b_0) + (a_1 \text{PrE}) + (b_0 [\text{Edu}] + (b_1 \text{PrE}) [\text{Edu}])$$

Without the parentheses and brackets, the variable PrE and Edu would appear on par in the regression equation; both would appear to be causal factors. The mysterious cross-term (b_1 PrE) [Edu] would call for interpretation. With this notation, the different roles of PrE and Edu are apparent. Edu is the only causal factor, while PrE affects the strength of the relationship, and is needed for quantifying the causal effect.

Focus on Mechanism: Currently, regression models search for invariance in the wrong place – they focus on trying to find ways to specify (A) and (B) such that these coefficients become constant. However, there is no reason or need that the strength of the causal effect should remain constant over large sets of environmental configurations. Although nothing remains constant over long stretches of time, it is the MECHANISM which remains stable over (short periods of) time. It is this mechanism which allows us to use current patterns to forecast future patterns, and also to assess effects of policy interventions. To illustrate how the search for mechanism shifts the focus of the research effort, think about HOW education affects earnings. One very simple factor comes to mind immediately. Some jobs require educational qualifications; you cannot apply for them without having the credentials. We will create a simple model to show how thinking about mechanisms leads to different ways to analyze the earnings education relationship.

A Simple Model for Earnings-Education Relationship: Suppose there are three sectors in the economy: Private, Government (Public), and Informal. Suppose that wages are $WP > WG > WI$; private sector offers highest wages, followed by government, while informal sector offers the lowest wages. Suppose that government sector jobs are restricted to the educated; application for these jobs requires a degree. Suppose that private sector does its own skills assessment and training, and therefore does not require degrees. Suppose that the educated prefer government jobs because these offer job stability, fringe benefits, retirement and health plans, which are substantially superior to the private sector, even though the wages are lower. Suppose there are 100 jobs in private and informal sector and 50 jobs in Government sector. Suppose there are 50 educated people in the population, who all get government jobs, and 200 uneducated, who split evenly between the private sector and the informal sector. If the average of WP and WI is above WG, then our Earnings education will show that education has a negative impact on earnings, even though it is actually quite beneficial. Attempts to find better fitting equations by adding variables and nonlinearities will be quite misleading and useless, because they do not reflect efforts to find out WHY more education leads to higher earnings. It is only focus on the mechanism, which remains stable under change, which will lead to improved understanding, and therefore improved estimates of the strength of the causal effects.

Failure of Positivist Paradigm: At the root of the problem is the positivist paradigm, which tell us to avoid thinking about the hidden causal mechanism of reality, and focus on the observables. Econometrics attempts to find stable relationships among the observables, but the only stable relationship occur at the level of causal mechanisms which underlie the observations. We will illustrate this issue with another few examples.

The Forecast Competitions M1 to M4: The International Journal of Forecasting organizes forecasting competitions as follows. In the most recent one (M4), 100,000 data series

were picked, and competitors were invited to submit forecasting methods. Methods which performed well overall in terms of average forecasting error over all the 100,000 series were awarded prizes in the competition. Forecast performance depends on the match of the model with the underlying real-world process which generates the data. If we forecast growth of a child, a S-shaped curve would do well. Forecasting oil prices would require attention to the dynamics of supply and demand of oil. If we just look at the numbers, success of forecasting is purely random – like using a crystal ball for forecasting. The results of the competition, when examined closely, validate this idea. Performance of forecasting methods varies widely across different series, and for different time periods. In any competition of random numbers, some will come out on top and others will come out on bottom. Nothing systematic can be learnt from such competitions. To go beyond crystal ball forecasting, one needs to look beyond the observed numbers to the real-world processes which generate the numbers.

Relating Advertising Expenses to Sales: As our last example, we consider assessing the relationship between Advertising Expense (Adv) and Sales (Sls). We have good a priori reasons, as well as strong empirical evidence, of the validity of the causal relationship $\text{Adv} \Rightarrow \text{Sls}$. However, the strength of the causal effect is of great importance to business firms trying to allocate budget to advertising. This means that we want to estimate A, B in the causal regression model $\text{Sls} \leq (A) + (B) [\text{Adv}]$. We have every reason to believe that the impact coefficient B will be affected by many different variables, when we look at the mechanism by which advertising leads to consumer decisions to buy. So there is no reason to expect to see a single stable value for B persist over a long period of time. The products, competitors and their advertising strategies, consumers, and the market, all change and evolve quite rapidly. Advertising strategies effective yesterday will be less effective in the rapidly changing world of social media. Running a regression to try to pick up a stable relationship is looking for an “invariant” relation in the coefficients, which does not exist. There is some invariance when we look to consumer decisions for purchase, which are based on perceived needs, information available from different sources, and perceived advantages and disadvantages of consumption choices. When we look at the stable mechanism, and then assess the impact of advertising on different aspects of this mechanism, we may be able to come up with advertising strategies which work under rapidly changing scenarios. Without examining the mechanisms, playing with sales and advertising numbers to try to find a good regression fit is a completely useless exercise.

Concluding Remarks: What insights does causality provide about regression models? Given a collection of variables, (X, Y, Z_1, \dots, Z_n) , our first task must be to search for DIRECT causal linkages – those which operate when all other variables are fixed. These are the AUTONOMOUS relations. Indirect chains are just sequences of direct chains. These are the CONFLUENT relations. Common Causes create APPEARANCE of relationship – these are SPURIOUS relations. Because there is no serious discussion or understanding of causality in most widely used econometrics textbooks, VAST amount of econometrics is study of SPURIOUS relationships. But, even when a genuine causal relationship is under study, econometrics searches for regression form which are “invariant” – that is, regressions which have stable coefficients. As we have seen, invariance occurs at the level of the underlying causal mechanism, and not at the level of the observed relationship. Thus, we endorse the Lucas critique

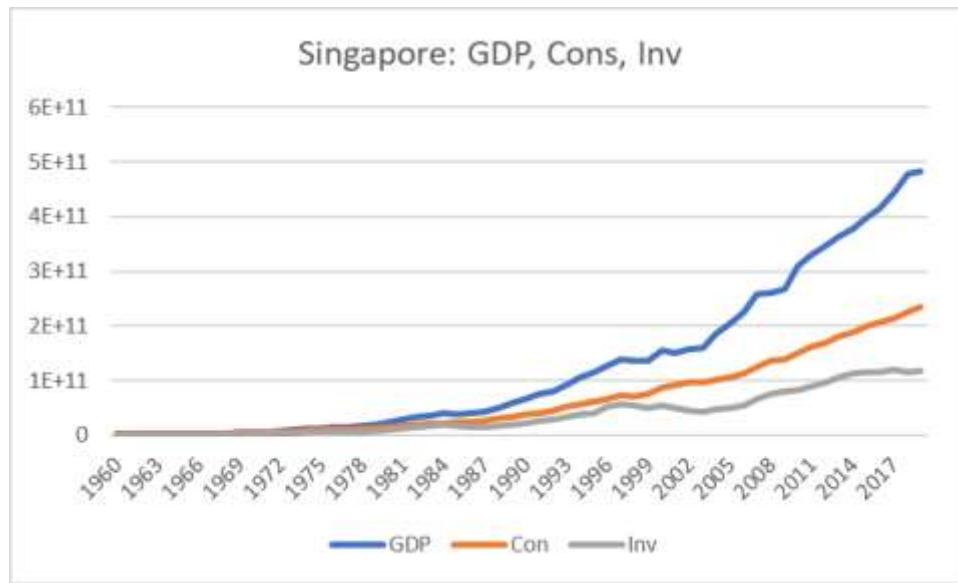
– regression failure occurs when we study surface relationships among observed variables, instead of searching out the underlying causal relationships. However, the search for causality cannot be done by axiomatic, *a priori* theory, as proposed by Lucas. Instead, we need an empirical approach to look at the real-world mechanism by which one variable affects another.

8E: Spurious Regressions

In previous lectures, we have seen the essential importance of analyzing the causal structure of the variables under study via path diagrams. In this lecture, we study one aspect of this which creates serious problems for understanding and interpreting regression results. This is the problem of a common cause. Z is a common cause of X and Y if $Z \Rightarrow X$ and $Z \Rightarrow Y$. When this happens, X and Y will be correlated, but neither variable causes the other. In such situations, a regression of Y on X will give results which show, according to standard methods for analysis, that X is a strong determinant of Y . This is called a “spurious” regression, or a “nonsense” regression. In this lecture, we will study some examples of this phenomena in real world data. Regressions for the lecture below were done using the RegressIt package for EXCEL - available from: <https://regressit.com/>

A Causal Relationship: A Consumption Function

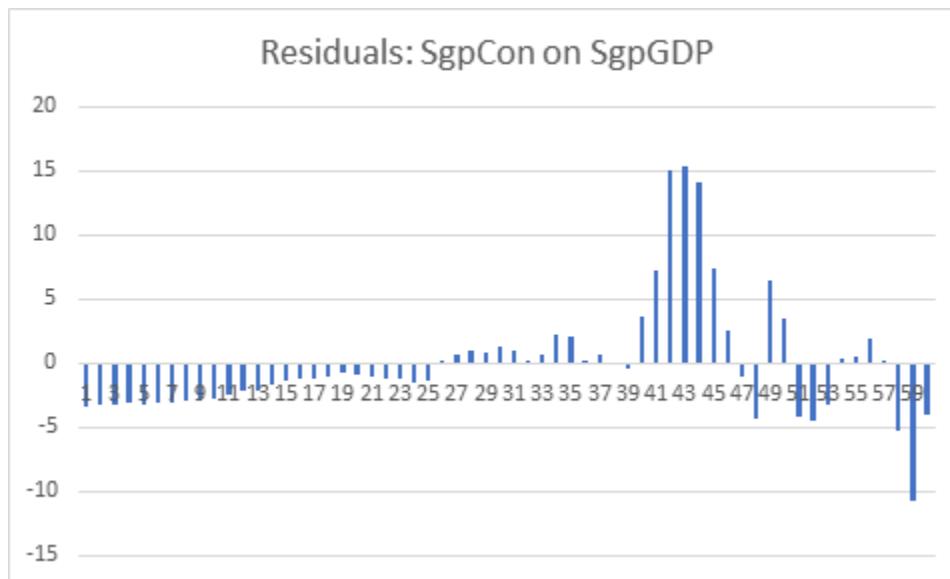
Data on annual GDP, Consumption, and Investment, for Singapore, taken from the WDI data set of the World Bank, is plotted below:



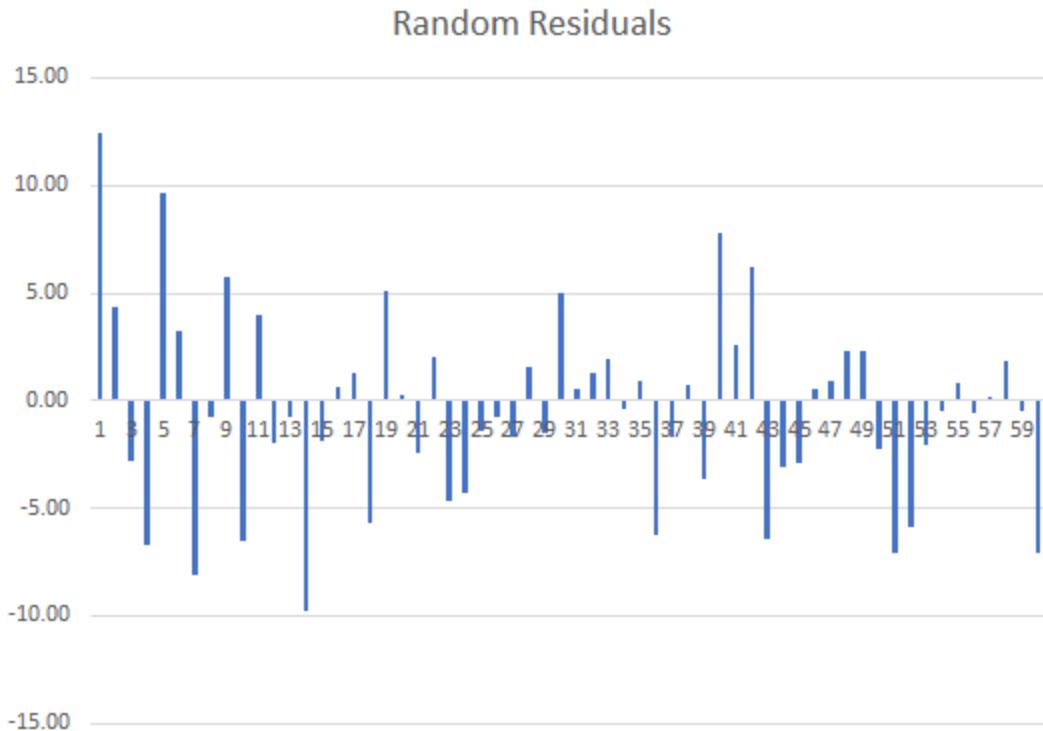
The Keynesian consumption function is one of the most widely accepted and estimated regression models. The causal hypothesis is that Income (GDP) determines Consumption (Con): $\text{GDP} \Rightarrow \text{Con}$. The simplest regression model which embodies this relationship is: $\text{Con} = a + b \text{ GDP}$. Running this regression on the data leads to the following results:

| Dependent Variable: | | SgpCon | | |
|----------------------------|-------------|------------|--------------|--------------|
| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var. |
| | 0.996 | 0.996 | 4.579 | 69.403 |
| Variable | Coefficient | Std.Err. | t-Statistic | P-value |
| Constant | 4.425 | 0.794 | 5.575 | 0.000 |
| SgpGDP | 0.486 | 0.004187 | 116.167 | 0.000 |

The regression has R-squared of 99.6%, which is interpreted to mean that 99.7% of the variation in Singapore Consumption can be explained by the Singapore GDP. The t-stat of 116 shows that the coefficient 0.486 of SgpGDP is highly significant. The p-value of 0.000 means we can reject the null hypothesis that the true coefficient is 0.0, corresponding to the idea that SgpGDP has no influence on SgpCon. Validity of regression results depends on a large number of assumptions, which are discussed in econometrics textbooks. One of the central assumptions is that the regression residuals should be random, and should come from a common distribution. To check whether or not this holds, we graph the regression residuals, the differences between the actual value and the regression fit:



This plot shows serious problems, since these residuals display systematic behavior. They are all negative and small early. To see how these patterns differ from independent random variables, we provide a graph of independent random variables with mean 0 and standard error 4.579, matching the estimated regression model standard error.



Random residuals frequently switch signs. They do not display any patterns in sequencing. The patterned residuals in the consumption function prove that the regression is not valid. In such situations, econometricians typically assume that the problem is due to missing regressors or wrong functional form. By adding suitable additional regressors, and modifying the functional form, one can generally ensure that the residuals appear to satisfy the assumptions made about them. But, solving the problem of random residuals by this search over regression models creates a serious problem. This can be explained as follows. Some of the fits in the data reflect a genuine real-world relationship. Other patterns are only accidental, and do not reflect any genuine relationship. The more we search, the more likely we are to end up with an accidental pattern. The fact that most regression relationships breakdown very frequently is due to the fact that most of them are accidental patterns in the data without any counterpart in reality.

The coefficient of the regressor is supposed to be a measure of the causal effect. According to the regression above, if SgpGDP goes up by \$100, then \$49 of it will be spent on consumption. This means that the enormously high proportion of 51% of the income will be saved. Since savings translate to investments, this could account for the dramatic growth of Singapore over the period in question. However, the patterns shown in the errors show that we cannot rely on the validity of this estimate. Vast amounts of experience with estimating this kind of regression function leads to two major conclusions.

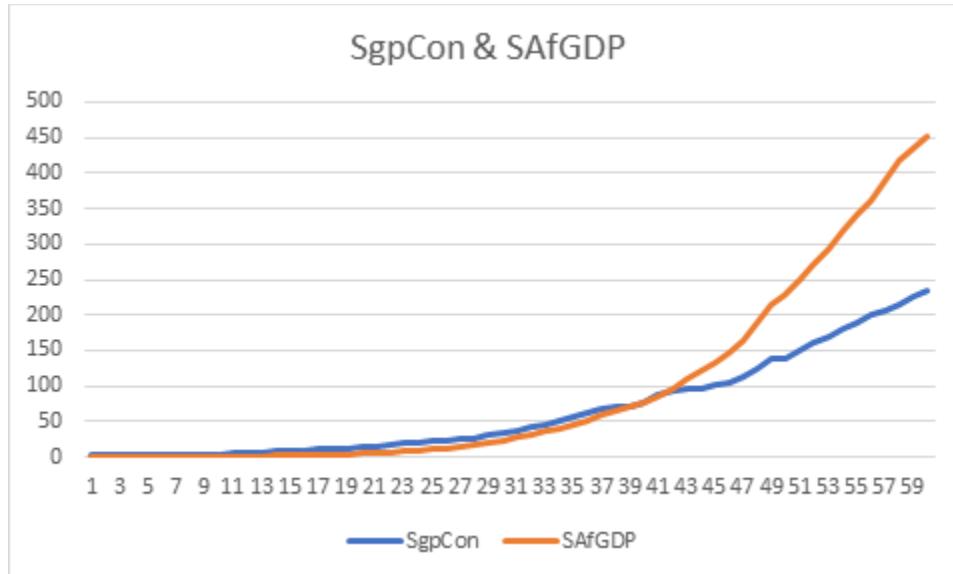
1. The estimate 0.49 of the causal effect is highly unstable – by adding many different plausible variables, we can make it change over a great range of values. Thus, these regressions do not provide accurate estimates of the key parameter we want to estimate.

2. The best fitting regressions very often make huge forecasting errors, and therefore are not very useful for policy.

Even though the regression is not very useful in estimating the size of the causal effect, at least the high R-squared SEEMS to confirm the causal effect that SgpGDP => SdpCon. At least, that was widely believed before the 1980's: high R-squared means that you have a good regression equation. It is this issue that we want to examine in this lecture. That is, we want to clarify how nonsense regressions can have high R-squared.

A Spurious Regression: Correlation without Causation

First, let us just demonstrate the problem by running a nonsense regression of SgpCon on SAfGDP – what is the relationship of Singapore Consumption with South African GDP. Before running the regression, we know that it is nonsense – there should be no relationship, or very little relationship between these two variables. Here is a graph of the two series, re-scaled:



A regression of SgpCon on SAfGDP yields: $\text{SgpCon} = 16.99 + 0.517 \text{ SAfGDP}$:

| Dependent Variable: | | SgpCon | | |
|----------------------------|--------------------|-------------------|---------------------|---------------------|
| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var. |
| | 0.966 | 0.966 | 12.848 | 69.403 |
| Variable | Coefficient | Std.Err. | t-Statistic | P-value |
| Constant | 16.994 | 2.048 | 8.299 | 0.000 |
| SAfGDP | 0.517 | 0.013 | 40.787 | 0.000 |

R-squared is 97%, and the t-stat of 40.8 on the coefficient 0.517 is very high. Based on the p-value of 0.000, we can safely reject the null hypothesis that the coefficient of SAfGDP is 0. But this is a nonsensical conclusion, if interpreted causally. It is not even remotely possible that a 100 Rand increase in South African GDP will increase consumption in Singapore by 51.7 Singapore Dollars. The central goal of this lecture is to highlight the dramatic difference between this regression and the previous one. While the previous equation of SgpCon on SgpGDP suffers from many defects, it is built on the right foundations of a causal relationship: SgpGDP => SgpCon. The estimated causal coefficient of 0.486 is most likely wrong, because the residuals show patterns violating the assumptions of the regression model. But there is a right coefficient, and we can have hope that we can get to a better estimate by adjusting the equation to fix the flaws in it. In contrast, there is no causal relationship between SAfGDP and SgpCon. The estimated coefficient 0.517 of SAfGDP in the second equation is purely mythical – it is measuring something which does not exist. What lesson do we learn from the fact that, despite this difference, both equations look more or less the same on the surface, with respect to the statistics produced by the regression package? This is discussed in the next section.

The Distinction Between Nominal and Real Econometrics

The central problem with conventional econometrics textbooks is the failure to clarify the difference between correlation and causation. The world of difference between the first and second equation above exists because the first equation attempts to estimate a causal coefficient, while the second provides an estimate of the correlation. Most students of conventional econometrics are never taught the difference between the two. To understand the deeper reason for this failure, it is useful to distinguish between “nominal” and “real” econometrics. By nominal, we mean econometrics techniques for which the names of the variables X and Y are sufficient, and we have no concern with the meaning of these names in the real world. The vast majority of econometrics taught in popular textbooks is nominal – it can be applied to any X and Y. In contrast, real econometrics is concerned with meaning. It is an aspect of real econometrics that the consumption function involves regression of Consumption on GDP and not the other way around. We know that the causation runs from income to consumption in the real world. ALL causal relationships are real world relationships which reflect the operation of real-world mechanisms linking the variables under study. Real relationships cannot be studied within a nominal framework, and this is why conventional textbooks fail to come to grips with the problems created by nonsense regressions. The distinction between the regression of SgpCon on SgpGDP versus SgpCon on SAfGDP is clear in real econometrics, but cannot be explained in nominal econometrics. More generally, conventional regression methods fail to specify the causal structures governing the variables under study, and hence fail to differentiate between causation and correlation.

It is useful to clarify the correct interpretation of the coefficient 0.517 of SAfGDP in the second regression. This measures a correlation that increases in SAfGDP and in SgpCon has happened together (correlation) in the past. If we observe a 100 Rand increase in SAfGDP, we could predict a \$51.7 in SgpCon, on the basis of past patterns of correlation. Because this

correlation is not based on any underlying causal relationship, it could easily break down in the future. Furthermore, one essential difference between causation and correlation has to do with INTERVENTIONS. The meaning of the causal relationship $X \Rightarrow Y$ is that changes in X will lead to changes in Y . In contrast, correlations break down under interventions. If we change SAfGDP, we would not expect to see any change in SgpCon. But we know this from “real” considerations – we cannot learn this from anything in the data by themselves.

The Nominalist Solutions:

The deeper point is that assessing whether or not a relationship is causal **ALWAYS** requires going beyond the names of the variables to the real-world concepts represented by the variables. This is why nominal econometrics is unable to distinguish between sensible regressions and nonsense regressions. This point will be argued and defended in detail later in the course. At this point, we will simply illustrate one class of nonsense regressions which have been widely recognized and acknowledged in the econometrics literature. Decades have been spent searching for a cure for this problem – how to distinguish between sense and nonsense – but no answers have been found. This is because no answers can be found within the methodology of nominalist econometrics.

The real solution to the problem of nonsense regressions can only be found by deeper study of causal structures connecting the variables under study. This involves stepping outside the boundaries of nominal econometrics. We will discuss some basics of real econometrics based on causal path diagrams later in this course. The phenomenon of spurious or nonnonsense regression was highlighted by Granger and Newbold in 1980, the search for solutions has spanned the past few decades. None of these have been successful. We discuss three main approaches developed for this problem below.

Missing Variables:

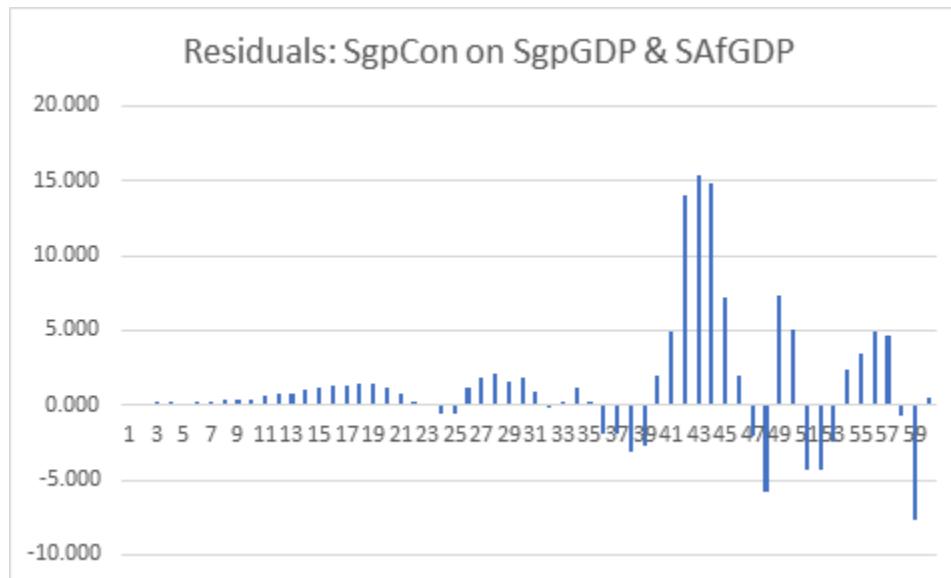
In the second regression, which regresses SgpCon on SAfGDP, the regression has a very high R-squared, and the coefficient of South African GDP is highly significant. This conveys the misleading message that SAfGDP is an extremely important determinant of SgpCon. The traditional “nominalist” understanding of this problem is that it arises from missing variables. The equation is giving us the wrong message, because the central determinant of SgpCon, which is SgpGDP, has been excluded from the equation. This leads to “Omitted Variables Bias”. Failure to put in the correct variables leads to SgpGDP acting as a proxy for the omitted variable, which is why it has a significant coefficient. Based on this approach, we can correct the problem by putting in the missing variable. Accordingly, we run a regression of SgpCon on both SgpGDP and SAfGDP. We expect that now, the SdgGDP will come out significant, and the SAfGDP will no longer be significant. However, the regression result is the following:

| | | |
|----------------------------|--------|--|
| Dependent Variable: | SgpCon | |
|----------------------------|--------|--|

| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var. |
|----------|-------------|------------|--------------|--------------|
| | 0.997 | 0.996 | 4.152 | 69.403 |
| Variable | Coefficient | Std.Err. | t-Statistic | P-value |
| Constant | 2.225 | 0.936 | 2.377 | 0.021 |
| SAfGDP | -0.103 | 0.028 | -3.681 | 0.001 |
| SgpGDP | 0.581 | 0.026 | 22.324 | 0.000 |

This time both variables – SAfGDP and SgpGDP – come out highly significant.

According to this regression, SAfGDP has a NEGATIVE effect on SgpCon. If South African GDP goes up by 100 Rand, SgpCon will decline by \$10.3. This is a nonsensical conclusion – but we can only say this because of our knowledge of the real-world causal relationships between the variables under study. Within the framework of nominal econometrics, it is impossible to explain the results of the above regression. However, looking at the residuals, which show patterns, we can reject this regression, and continue our search for a better model:



This shows the weakness of the missing variables strategy. There are too many variables to try, and no way to know when you have hit the right combination. By experimentation with variables and functional forms, econometricians often succeed in getting a regression equation which matches our intuitive understanding of the causal effects relating the variables under study, and also has residuals which look random. But this is due to the skill and experimentation done by the econometrician, and does not reflect information conveyed by the data itself.

Common Time Trends:

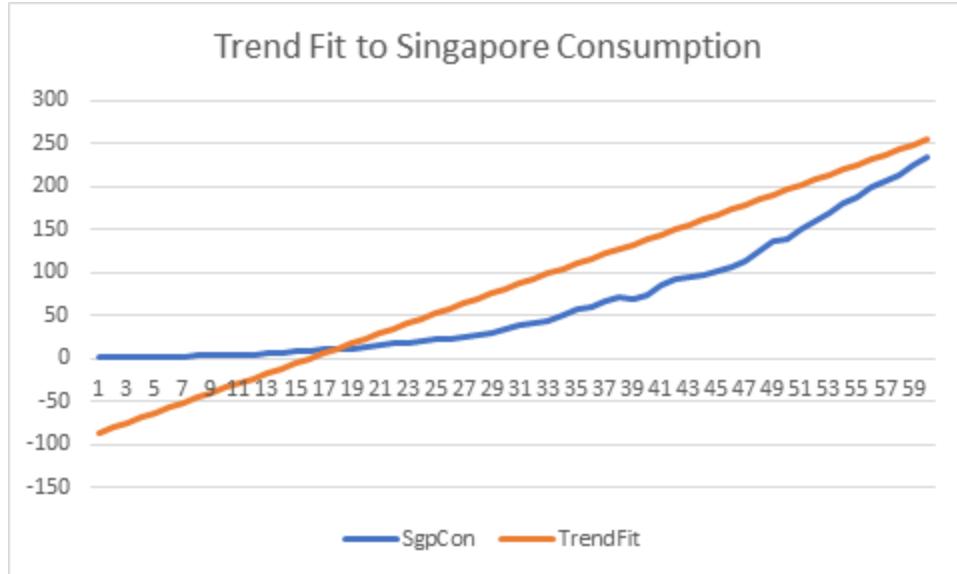
A second approach to the spurious regression of SgpCon on SAfGDP comes from noting that both variables have strong time trends. We can think of this as a common factor driving both variables. T => SdpCon and T => SAfGDP. Variations in the common factor cause both variables to vary together, and create a misleading impression of a causal relationship between the two. One way to handle this problem is called “detrending”. This involves removing the effect of time trend from both variables. What is left, after removing the trend, should not suffer from the problem of spurious correlation. Detrending is done by running a regression of a variable on a suitable function of Time, and then taking the residuals from this regression to be the detrended variable. We illustrate how this works.

A second approach to the spurious regression of SgpCon on SAfGDP comes from noting that both variables have strong time trends. We can think of this as a common factor driving both variables. T => SdpCon and T => SAfGDP. Variations in the common factor cause both variables to vary together, and create a misleading impression of a causal relationship between the two. One way to handle this problem is called “detrending”. This involves removing the effect of time trend from both variables. What is left, after removing the trend, should not suffer from the problem of spurious correlation. Detrending is done by running a regression of a variable on a suitable function of Time, and then taking the residuals from this regression to be the detrended variable. We illustrate how this works.

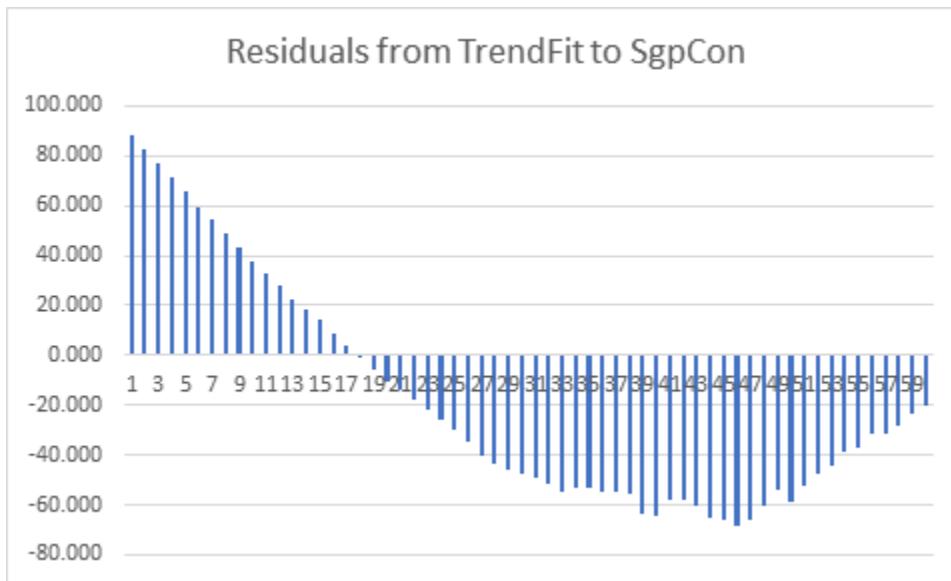
If we fit a simple linear trend to SgpCon, we get $SgpCon = -91.742 + 5.775 T$:

| Dependent Variable: | | SAfCon | | |
|----------------------------|-------------|------------|--------------|--------------|
| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var. |
| | 0.721 | 0.716 | 63.238 | 118.752 |
| Variable | Coefficient | Std.Err. | t-Statistic | P-value |
| Constant | -91.742 | 16.534 | -5.549 | 0.000 |
| Trend | 5.775 | 0.471 | 12.250 | 0.000 |

We can plot the actual data against the fitted time trend as follows:



The goal of fitting a trend to the data is to “detrend” it – that is, to make the residuals independent of time. However, the detrended residuals from the above regression can be plotted as follows:



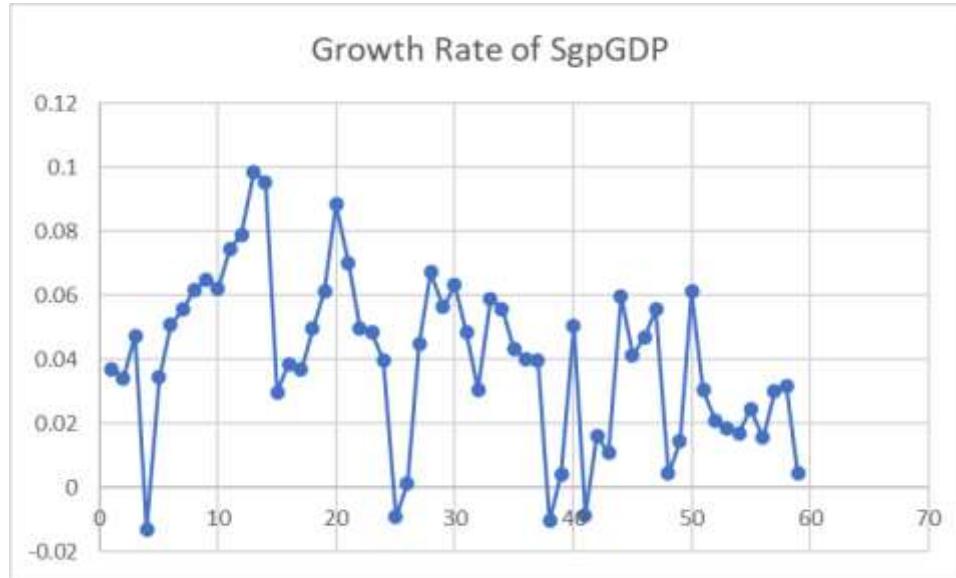
These residuals are definitely not independent of time – they show a very strong pattern over time, and look very far from the random residuals required by regression assumptions.

Staying within the framework of nominal econometrics, it is possible to solve this problem by using more complicated time trend functions. For example, if we add a squared term T^2 , this would substantially improve the fit, and remove much of the pattern that we see in the regressors. By increasing the degree of the polynomial, we can eventually get an excellent fit and ensure that the residuals have no remaining correlation with time. However, this mechanical

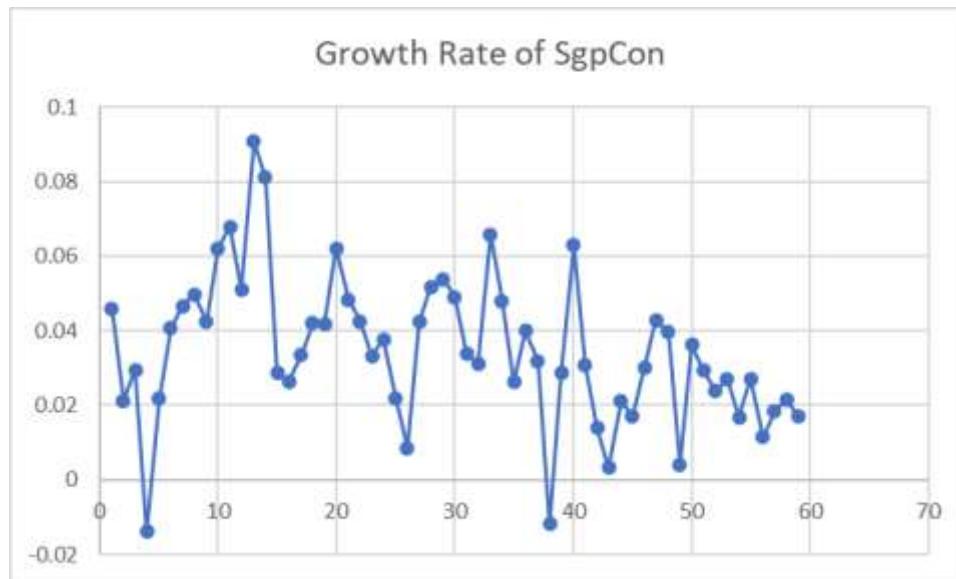
strategy of nominal econometrics is not very useful. This is because the higher degree polynomials do not represent the real-world mechanism which creates the growth process in the GDP. So we can conclude that, like the missing variables strategy, the detrending strategy fails to solve the problem of spurious regressions in this case.

Variable Transformation to Stationarity:

The reason for the failure of trend fit may be due to the reasonable assumption that the growth rate is stable across time, so that $\text{Gr}(\text{SgpG}) = \log (\text{SgpGDP}(t)/\text{SgpGDP}(t-1))$ would be uncorrelated with time. This can be checked by graphing this variable against time, as follows:



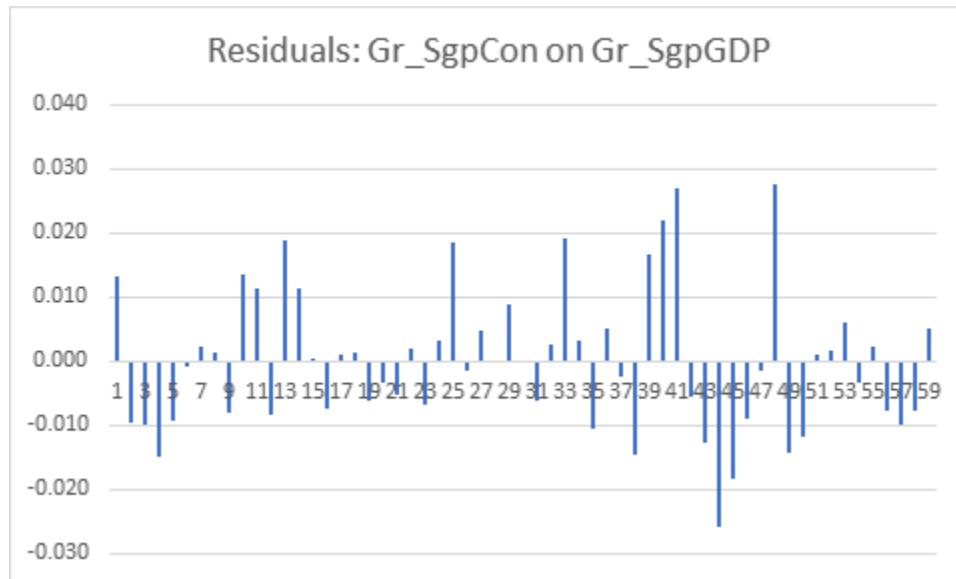
Now it seems clear that this variable does not show any time trend. We can similarly transform SgpCon to growth rates by defining $\text{Gr}(\text{SgpC}(t)) = \log(\text{SgpCon}(t)/\text{SgpCon}(t-1))$. Graphing this variable gives:



This variable is also uncorrelated with time, at least visually. The variable transformation of taking growth rates has removed the common time trend from both variables. Now it should be possible to run a regression of Consumption on Income which does not suffer from the spurious correlation between the two variables created by the common trend. Unfortunately, this strategy also fails to differentiate between genuine and spurious regressions as we will see. First, let us look at the genuine causal regression of Gr_SgpC on Gr_SgpG: $Gr_SdpC = 0.009 + 0.634 Gr_SgpG$

| Dependent Variable: | | Gr_SgpC | | |
|----------------------------|--------------------|-------------------|---------------------|---------------------|
| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var. |
| | 0.684 | 0.679 | 0.011 | 0.020 |
| Variable | Coefficient | Std.Err. | t-Statistic | P-value |
| Constant | 0.009254 | 0.002720 | 3.402 | 0.001 |
| Gr_SgpG | 0.634 | 0.057 | 11.119 | 0.000 |

R-squared is now 68%. It has gone down because we have removed the correlation due to the common factor of time-trend, but it is still very high. The regression provides a good fit. Also, for the first time, the residuals from this regression look random:

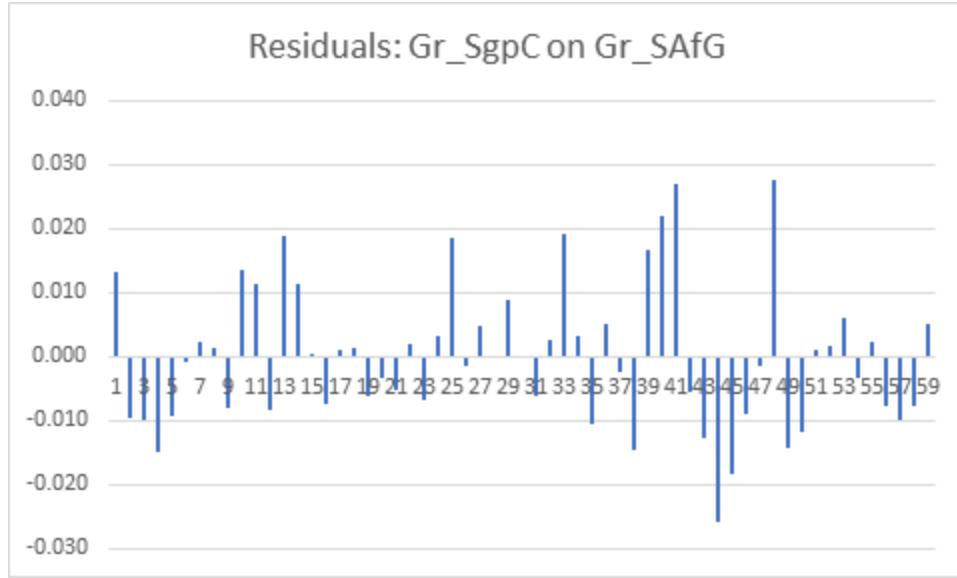


This transformation from levels to growth rates has substantially improved the regression results. There is now some hope that the coefficient 0.634 might be a correct measure of a causal effect, if other aspects of the specification are also correct.

However, our main concern is to see if this transformation can distinguish between genuine causal relationships and spurious ones. For this purpose, we run the regression of the growth rate of SgpCon on the growth rate of SAfGDP. This yields the following results:

| | | | | |
|---|-------------------|------------|--------------|-------------|
| Dependent Variable: | C | Gr_Sgp | | |
| Independent Variables: | | | | |
| Gr_SAfg | | | | |
| Equation : | | | | |
| Predicted Gr_SgpC = 0.006484 + 0.566*Gr_SAfg | | | | |
| <u>Regression Statistics: Model 6 for Gr_SgpC (1 variable, n=59)</u> | | | | |
| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var |
| | 0.308 | 0.296 | 0.017 | 0.020 |
| <u>Coefficient Estimates: Model 6 for Gr_SgpC (1 variable, n=59)</u> | | | | |
| Variable | Coefficien | | | |
| | t | Std.Err. | t-Statistic | P-value |
| Constant | 0.006484 | 8 | 0.00601 | 1.077 |
| Gr_SAfg | 0.566 | | 0.112 | 5.035 |
| | | | | 0.286 |
| | | | | 0.000 |

This regression also shows that growth rates of South African GDP are highly significant in explaining growth rates of Singaporean Consumption. The regression residuals look random, and support the validity of this regression:



According to conventional econometrics, the causal regression of growth rates of SgpGDP on growth rates of SgpCon allows us to conclude that 63% of changes in Gr_SgpGDP are transmitted to Gr_SgpCon. But, by exactly the same reasoning, we can use the spurious regression of Gr_SgpCon on Gr_SAfGDP to conclude that 56.6% of changes in the growth rate of the South African economy are transmitted to growth rates of Singaporean consumption. The first measurement has some chance being correct, while the second statement is ridiculous.

In this particular case, the missing variables strategy can solve the problem. If we think that the source of the problem in this last equation is that the primary determinant of SgpCon is missing from the equation, we can fix the problem by adding this variable. When we do so, we get the following equation:

| Dependent Variable: | | Gr_SgpC | | |
|----------------------------|-------------|------------|--------------|--------------|
| | R-Squared | Adj.R-Sqr. | Std.Err.Reg. | Std.Dep.Var. |
| Variable | Coefficient | Std.Err. | t-Statistic | P-value |
| Constant | 0.003740 | 0.003992 | 0.937 | 0.353 |
| Gr_SAfG | 0.163 | 0.088 | 1.855 | 0.069 |
| Gr_SgpG | 0.569 | 0.066 | 8.625 | 0.000 |

Putting in growth rates of both Singapore and South Africa leads to the right result: Singapore GDP is a highly significant determinant, while South African GDP is not significant at

the 95% level. Nominal econometrics consists of this kind of exercise, where we try one equation after another in order to get a match to our a priori ideas about the size and strength of the causal effects. This is quite the opposite of what students are taught about this methodology. It is not that we allow the data to tell us about the causal structures in the world. Rather, we know these structures in advance, and try many different formulations until we get one which matches our preconceptions.

Concluding Remarks:

Econometricians have been working on finding methods to discriminate between nonsense and sensible regressions for decades, without success. We have discussed three strategies for doing so in this lecture. All three lead to failure, although the third one can be salvaged. Currently, it is the third strategy which dominates the scene. However, it also suffers from failures in many different cases. The reason for this failure is that solutions to causal problems cannot be found in nominal econometrics. Without having a good grasp of the causal structures which relate the regressors to the dependent variables, it is not possible to estimate causal effects.

EXCEL SPREADSHEET with data & regressions: [SgpCon2Models.xlsx](#) - Regressions for the lecture below were done using the RegressIt package for EXCEL - available from: <https://regressit.com/>

SLIDES for the Video-Talk: [RSIA10E Spurious.pptx](#)

9: Assessing Association Between Two Series

Here we must put in a BLURB about what the central issues treated in this chapter are.

9A Exports & Growth Part 1

1 Association, Correlation, and Causality

An important goal of Real Statistics is the discovery of Causal Relationships. Note that Causality is never observable – it cannot be revealed by the data. Suppose X is a Cause of Y: this means that “If X had been different, this would have affected Y”. This imaginary event is never observable. This was noted by philosopher Hume, who said we can only observe X followed by Y and never observe that X CAUSES Y. Nonetheless, discovery of causation is of greatest importance. It requires learning how to use the patterns among the observables to learn about the unobservable real world factors which govern the observables. A major cause of the blindness to causality was the emergence of the philosophy of logical positivism, patterned after the thoughts of Hume. Logical Positivists claimed that science must concern itself solely with observables. This idea exercised an enormous influence on shaping the foundations of the social sciences in the early 20th Century. It was responsible for the widespread view that “causality” was a meaningless concept, because was not observable in the data.

In this lecture, we will only study the “association” between two variables, which is very much an observable concept. It is important to note that associations are SYMMETRIC, while causality is asymmetric. The usual name for “association” in statistics is correlation, but we will avoid that term for reasons to be clarified later. Statisticians understand the difference between the observable correlation and the unobservable causation in the popular maxim: “Correlation is not Causation”. However, they restrict study to correlations, and ignore causation.

We will see that associations can be used to distinguish between many different kinds of causal structures. Association provide CLUES to underlying causal structures. Further exploration of real world may confirm or reject these clues. Even though discovery of Causal Structures is of utmost importance, unfortunate historical events led to the banishment of causality from statistics, until recently. Pearl et. al. provide the detailed story in Chapter 2 of The Book of Why about How Statistics became blind to causality. They describe how positivist philosophy led one of the founding fathers of statistics, Karl Pearson to think that: “*Meaningful thoughts can only reflect patterns of observations, and these can be completely described by correlations. Having decided that correlation was a more universal descriptor of human thought than causation, Pearson was prepared to discard causation completely.*”

2 Export Led Growth

In this lecture, we will study the World Bank WDI data set on exports and growth to assess if these two variables are “associated” with each other. We will discuss the observable patterns in the data regarding the two variables and use these patterns to assess the strength of different hypotheses about the causal relationships between the two variables. Literature on the topic provides support for ALL four possibilities listed below.

1. Export-Led Growth (ELG): This is the idea that ‘Exports are drivers of growth’. Thus, to achieve high growth, nations should focus in increasing exports.
2. The anti-thesis of ELG is Growth Led Export. When an economy grows, it will naturally export more, creating an association between exports and growth which can be misinterpreted.
3. A third possibility is that of “Mutual Causation”. As an economy grows, exports will increase, creating linkages between domestic and global economy. These linkages will further promote growth creating a virtuous cycle where exports and growth grow together and reinforce each other.
4. The fourth possibility is that there is no relationship between the two. Growth and exports are driven by different forces and factors, and any observed relationship between the two is purely coincidental.

We would like to examine the WDI data set to see what it tells us about these four hypotheses. The first step is to pick out the relevant variables. There are a large number of different series in the WDI data set both about exports and GDP, which could be used to examine the relationships. After looking through many different series, the following two seem to most

directly relevant to the question under discussion. The material below is taken directly from WDI Data set descriptions of variables.

GNI growth rates: NY.GNP.MKTP.KD.ZG: Economic Policy & Debt: National accounts: Growth rates. GNI growth (annual %) - GNI (formerly GNP) is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad.

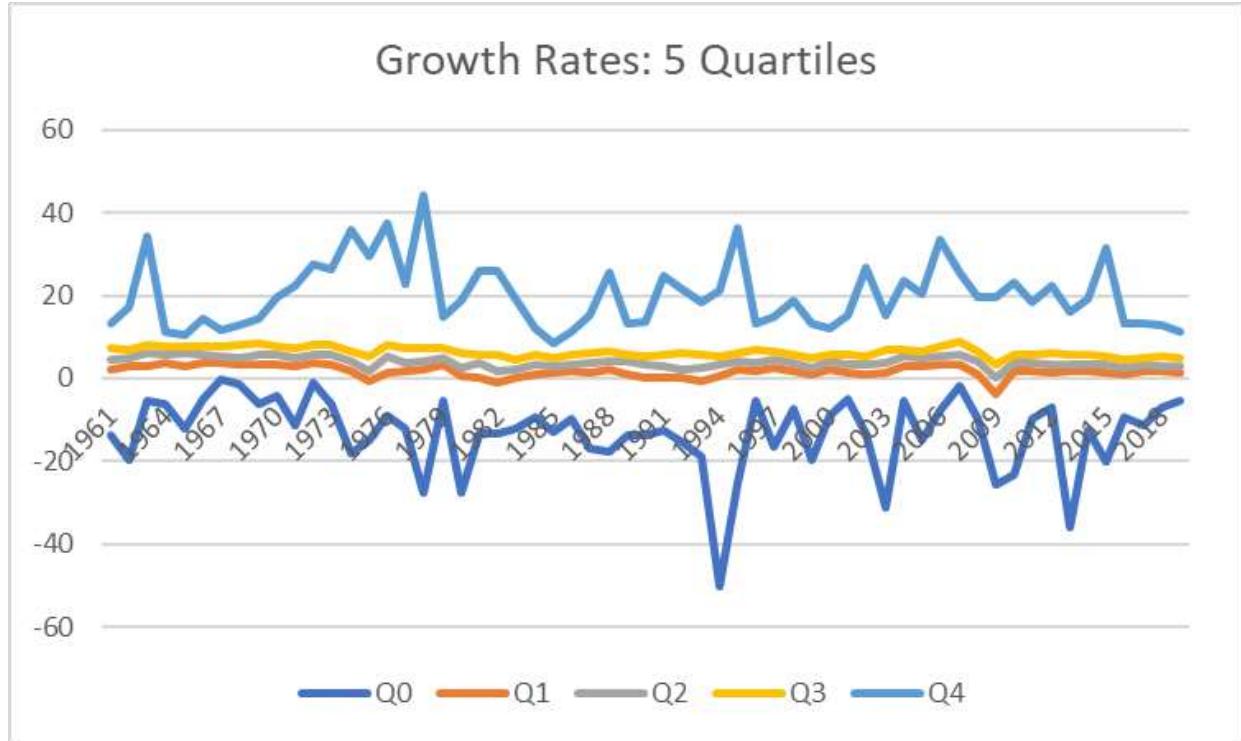
Ratio of Exports to Gross Domestic Product: NE.EXP.GNFS.ZS: Economic Policy & Debt: National accounts: Shares of GDP & other -Exports of goods and services (% of GDP) - Exports of goods and services represent the value of all goods and other market services provided to the rest of the world. They include the value of merchandise, freight, insurance, transport, travel, royalties, license fees, and other services, such as communication, construction, financial, information, business, personal, and government services. They exclude compensation of employees and investment income (formerly called factor services) and transfer payments.

There are a number of issues regarding how the data is collected, and what it means, but we will ignore these for the moment. The most important is these is the large amount of missing data for GNI growth rates. Even though the WDI data set has a nearly complete data set for GNI (GDP), it has a lot of missing data in the GNI growth rate series. This is a mystery because we can compute the growth rate from the GDP series. My guess is that a lot of the data points are guesstimates, which means that computing the rate of growth is like taking a ratio of GUESS 1 to GUESS 2, which will produce random numbers. So the data given is, somehow, on more solid grounds then the guesstimates for GDP.

3 Median Growth Rates Across Time

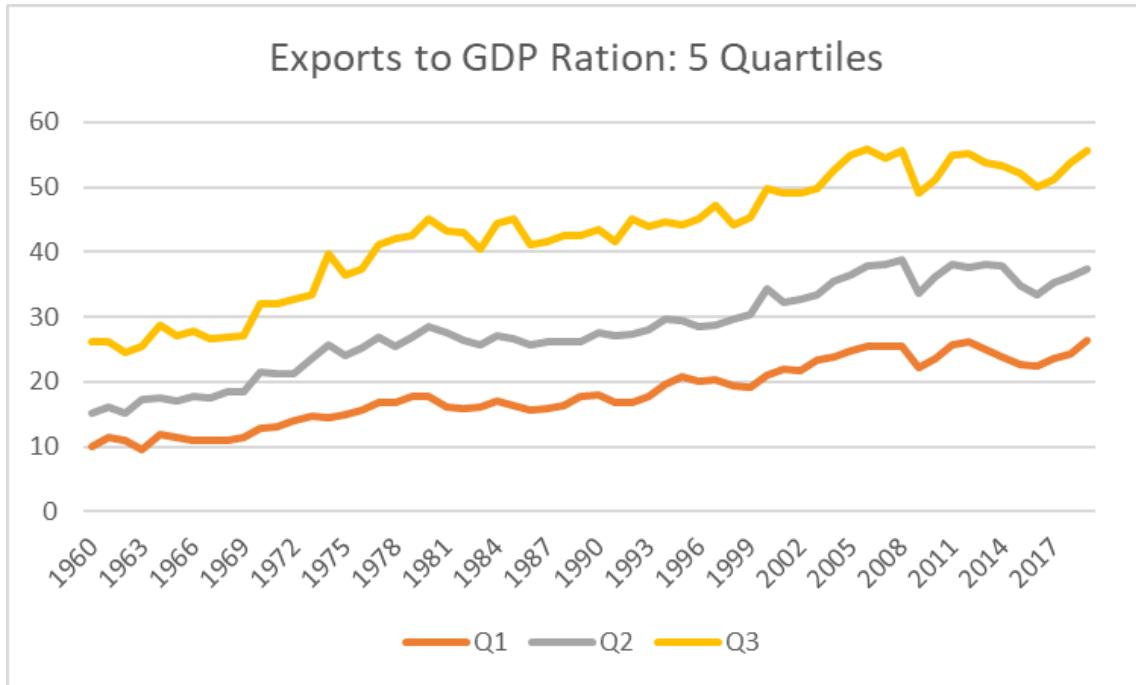
The Question we want to tackle in this lecture is: Are these two variables ASSOCIATED? Do the two move together? We will develop intuitive methods to assess this. Before we look at the two together, we look at both of the two variables separately, to get some idea of how they have changed over time.

The Five Quartiles of the GNI growth rates



As often happens, the central 3 quartiles are dwarfed by the Minimum (Q0) and the Maximum (Q4). Very low – highly negative – rates of growth are seen following some natural catastrophe like floods, tsunamis, etc. Also important are man-made catastrophes like wars, and economic crises of various kinds. Recovery following such disasters can sometimes be very rapid, and lead to very high growth rates. Although it is not relevant to our current topic, it would be worth investigating the highest growth rates achieved to learn more about the reasons for spectacularly good performance.

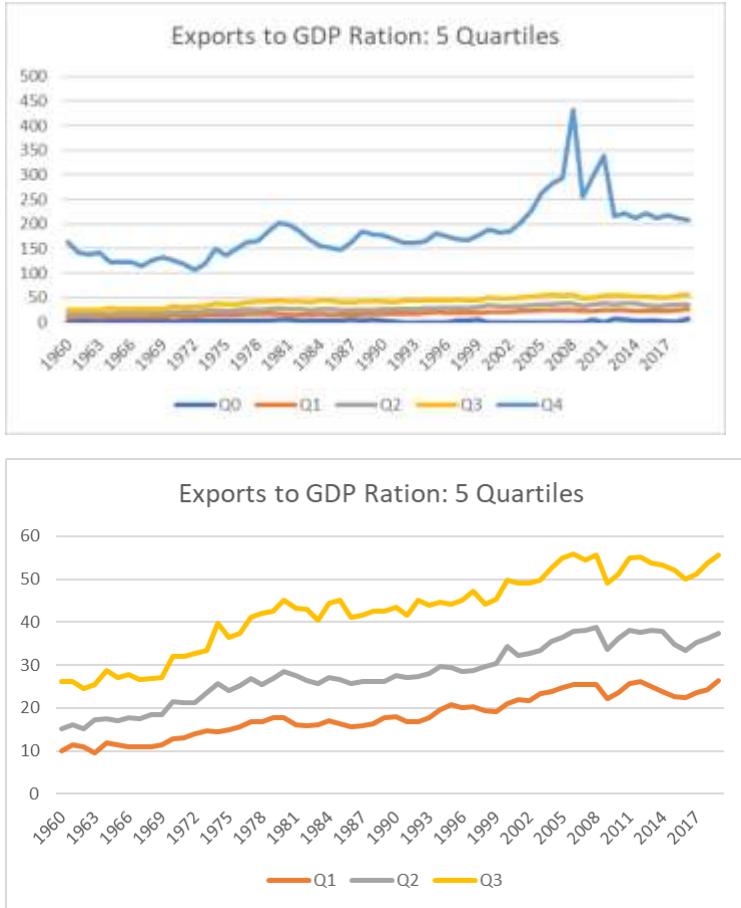
In order to get a clearer picture of the World Growth Rates, we focus on the middle 3 Quartiles:



Median growth rates were in the range of 4% to 6% until around 1973. Very few people realize the enormous impact of [the 1973 Oil Crisis](#) on the world we live in. Arab Oil Embargo in response to US support of Israel in the Yom-Kippur war led to tripling of oil prices, and created a recession in economies throughout the world. The steep drop in growth rates globally can be seen in the graph above. This crisis creates the opportunity for Chicago School free market economists to repudiate and discredit Keynesian economics, and re-install neoclassical economics on new mathematicised foundations. The Reagan-Thatcher era of free market economics started in the 1980's and saw depressed growth rates between 2% and 4% together with reduction or elimination of business cycles. This is known as the era of "The Great Moderation". However, removal of Depression era regulations on banking and finance in 1999 and 2000 led to a financial boom starting around 2003, which was followed by a spectacular crash in 2007. All of this real-world history is reflected in the global growth graphs above, and these graphs cannot be understood without an understanding of this historical background about the real world events which generated these graphs.

4 Median Export to GDP ratio Across Time

Next we look at the ratio of Exports to GDP across time: The 5 quartiles are given in the following graph:



Since this ratio can never be

negative, the minimum remains close to zero and does not distort the scale of the graph. However, the maximum is huge, and does distort the scale. It does show an interesting pattern of dramatically high growth right after 2000, followed by crash in 2007. An unexpected pickup between 2008 and 2011, followed by a return to normal levels. These are clues worth following up, but not relevant to our present research question. To get a clearer picture, we plot the three quartiles separately as follows:

The median export to GDP ratio over all countries shows a rising trend from 1960 to 1980, going from 15% to 30% over this period of time. From 1980 to around 1999 this ratio remains stagnant, slightly under 30%. Then there is period of take-off, where exports grow to around 40% before crashing in 2007, but still remaining above 30%. There is a lot of history of international trade reflected in the graph, but we will bypass this to proceed to discuss our main question.

9B: Failure of Exports-Led Growth Hypothesis

Ranking Countries According to Growth and Exports

To find out if growth rates and exports are associated, we would like to see if high growth countries also have high exports and vice versa. Similarly, we can check to see if low growth countries have low exports and vice versa. The first step is to CLASSIFY countries as being high or low growth. This is a problem because we have many years of data – a country which is very high growth in some years may have low growth on other years. For example, consider the growth rates of Turkey and Thailand

| | 1 993 | 1 994 | 1 995 | 1 996 | 1 997 | 1 998 | 1 999 | 1 000 | 2 001 | 2 |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|---|
| Turkey | 7 .8 | - 5.5 | 8 .5 | 7 .6 | 7 .6 | 2 .8 | - 3.7 | - .6 | 6 7.0 | - |
| Thailand | 8 .9 | 7 .9 | 7 .9 | 5 .0 | - 3.2 | - 8.5 | 5 .6 | 5 .7 | 5 .6 | 2 |

Comparison of Turkey and Thailand on Growth Rates.

In 93,94,99, and '01 Thailand had higher growth rates than Turkey. In the other years, Turkey had higher growth rates. If we are to classify countries in terms of growth rates, which of the two should rank higher? We have already studied this problem in detail earlier in this course. We know that there is no objective way of ranking countries when comparisons are being made on multiple dimensions. If we give more weight to recent years, we will come up with one criterion. If we give equal weight to all years, this will give a different result. There are many ways that we can come up with a single number ranking, none of which are objective.

The purpose of ranking here is to see if countries with higher growth also have higher exports. Our ranking does not have to perfect for this purpose. As long as the ranking is roughly aligned with growth rates, and differentiates between high growth rate countries and low growth rate countries, it will be adequate for our purpose. We propose the use of the Median growth rate as the single number to rank countries. In the above example, Turkey has median growth rate of 6.6% – 4 of the growth rates are below this value, and 4 of the growth rates are above this value in the 9 years of data. Similarly, Thailand has median growth rate of 5.6% – 4 rates are above and 4 are below. On the basis of median growth rates, Turkey (6.6%) ranks higher than Thailand (5.6%) for these 9 years of data. We note that on the full data set this ranking is reversed, and Thailand ranks slightly higher than Turkey.

Having decided on the criterion to use for ranking countries according to growth, it is an easy matter to calculate. The data for growth rates goes from column E to BL corresponding to years 1960 and 2019. The countries are listed starting from row 4. In the column BM, write Median Growth, and enter =MEDIAN(E4:BL4) in the 4th row to compute the median growth rate of the data for the country in the 4th row. Copy this down the column to compute the median growth for all countries in the data set. One very important fact to note is that EXCEL ignores missing data in this calculation. Many of the countries have lots of missing data. EXCEL computes the median for the years for which data is available, ignoring the years for which data is not available. (This give a very different result from an alternative strategy, which replaces

missing data by 0's, for example.) Once we have calculate ONE NUMBER as the growth rate performance measure, it is easy to ask EXCEL to sort all the countries according to the column BM, the median growth rate for each country. This arranges the countries in decreasing order of (median) growth rates.

MISSING DATA: Interestingly, the 47 countries from row 175 to 221 have values of #NUM! – EXCEL cannot compute the median growth rate. When we look at the data set, we see that ALL growth rate values are missing for these countries. When all the data is missing, the median cannot be computed and EXCEL gives an error: #NUM!. This is very surprising because the WDI data set does not have much missing data for the GNI (or GDP). Once the GDP data is present, we can directly compute the growth rates from this data, to fill in the missing values for the growth rates. My speculation is the World Bank uses guesswork to compute the GDP for countries and years for which there is missing data. To compute the growth rates from these guesses at GDP values would substantially compound the errors by taking a ratio of two guesses.

We follow exactly the same procedure to rank countries in terms of the Exports to GDP ratio. Since these ratios fluctuate a great deal over time, there is no simple to assess which country has greater percentage exports. We use the median of all available data as the one number to measure export performance for each country. Once all of the data on the Export/GDP ratio is reduced to one number by taking the median of all available numbers, it becomes possible to rank the countries according to this number. The results from this ranking allow us to identify the top 20 – the best performers, and also the bottom 20 – the worst performers. We compare these best and worst according to both the criteria below.

This table below given the top 20 countries in terms of Median Growth Rates, and also the top 20 in terms of the Exports to GDP ratio. The highest ranked 20 countries with respect to median growth range from 6.5% to 9.4%, while to top 20 export to GDP ratios vary from 181% to 201%. Again, the important thing to note is that there is no overlap. The fastest growing countries are not the top exporters, and the top exporters are not the fastest growing countries. It is surprising that Ethiopia has the highest median growth rate among all countries in the data set. This is partly due to a data problem, but also partly due to the stellar performance of Ethiopia in the recent years, as clarified in the note on Ethiopia below.

NOTE on Ethiopia: It is worth clarifying the oddity of Ethiopia emerging as the top performer with respect to median growth. This is because the WDI data set only records growth rates from 2012 to 2018 for Ethiopia. If the data set had been full, Ethiopia would have ranked much lower, because it undoubtedly had a very poor growth rate over most this period from 1960 onwards. Ethiopia is a star performer only because of the missing data. Normally, data anomalies like this leads to outliers which must be ignored for correct analysis. In this case however, Ethiopia did genuinely have star performance, ranking close to the top in growth rates in the whole world, for six consecutive years. There are many articles on the Ethiopian growth miracle. It is clear the exports were not part of this miracle performance, so we learn that getting very high growth rates definitely does not depend on exports. That is, exports may cause growth, but lack of exports cannot be a barrier to growth.

Author Last Name/Book Title

| Name | Gr | med rank | Gr | Name | X/G | Med | X/G |
|---------------|----|----------|----|------|-------------|--------|-----|
| | | | | | | | rnk |
| Ethiopia | | 9.39% | | 171 | Virgin Isl. | 207.7% | 201 |
| China | | 8.62% | | 170 | San Marino | 173.0% | 200 |
| Qatar | | 7.93% | | 169 | Singapore | 168.9% | 199 |
| Macao | | 7.85% | | 168 | Djibouti | 149.2% | 198 |
| Singapore | | 7.85% | | 167 | Hong Kong | 113.9% | 197 |
| Korea, Rep. | | 7.67% | | 166 | S Maarten | 113.3% | 196 |
| Tajikistan | | 7.66% | | 165 | Luxembourg | 104.5% | 195 |
| Djibouti | | 7.60% | | 164 | Malta | 94.9% | 194 |
| Lao PDR | | 7.14% | | 163 | Macao | 86.4% | 193 |
| Kiribati | | 6.97% | | 162 | Bahrain | 83.9% | 192 |
| Azerbaijan | | 6.89% | | 161 | U A E | 82.2% | 191 |
| Malaysia | | 6.89% | | 160 | Maldives | 77.7% | 190 |
| Cote d'Ivoire | | 6.82% | | 159 | Am Samoa | 74.8% | 189 |
| Cambodia | | 6.80% | | 158 | Eq Guinea | 71.9% | 188 |
| Armenia | | 6.71% | | 157 | Ireland | 70.7% | 187 |
| Greenland | | 6.71% | | 156 | Aruba | 70.5% | 186 |
| Ghana | | 6.57% | | 155 | Estonia | 69.2% | 185 |
| Angola | | 6.55% | | 154 | Slovak Rep | 68.5% | 184 |
| Botswana | | 6.54% | | 153 | Brunei | 68.0% | 183 |
| Mozambique | | 6.53% | | 152 | Hungary | 66.9% | 182 |
| Bhutan | | 6.51% | | 151 | Malaysia | 66.6% | 181 |

Top 20 countries for growth rates compared with Top 20 Exporters (Highest Exp/GDP ratio)

After looking at the top 20 countries with respect to the median growth and exports criteria, we now look at the bottom 20. The median growth rates for these worst performing countries are between 0.3% and 2.2%. Similarly the bottom ranked countries for exports have export/GDP ratios varying from 0.47% to 12.8%. What is of great interest is that there is no overlap between these two sets of countries. The worst performers with respect to growth are NOT in the bottom 20 with respect to exports. Similarly, the countries with the lowest export to GDP ratios are not in the bottom 20 with respect to GDP growth. This suggests that the connection between exports and GDP growth is not very tight.

NOTE on Percentile Ranks: We will do a side-by-side comparison of growth and export performances of the fastest growing countries. To do this, it is useful to look at the RANKS of the countries, instead of the growth rates and the export/GDP ratios. The smallest growth rate (the worst performance) has the lowest rank, while the highest growth rate has the highest rank. In the entire data set of 218 countries, there is no data at all on 47 countries. These must be excluded from the data set, which leaves 171 countries for which we have data on growth. These are ranked from 1 to 171 in terms of increasing growth. Jamaica has the lowest rank of 1 and the smallest median growth rate of 0.3%, while Ethiopia has the highest rank of 171, with median growth rate of 9.4%. For Export to GDP ratio, data exists for 201 countries. Bottom ranked is Myanmar with Exp/GDP ratio of only 0.5% while top ranked Virgin Islands has Exp/GDP ratio of 207%. The Exp/GDP ranks vary from 1 to 201. To make the ranks comparable across the two categories, it is convenient to divide the rank number by the total number of countries in the data set – this is called the percentile rank. Then Ethiopia gets the percentile rank of $171/171 = 100\%$ showing that it is the top-ranked country for median growth rate. Similarly, Virgin islands gets the rank of $201/201 = 100\%$ showing that this is the top ranked country for Exp/GDP ratio.

| Country | Med gr | Gr rank | Country | X2G | Med rnk | X2G |
|--------------|--------|---------|--------------|-------|---------|-----|
| Jamaica | 0.31% | 1 | Myanmar | 0.47% | 1 | |
| Palau | 0.49% | 2 | Fr Polynesia | 3.88% | 2 | |
| Sint Maarten | 0.74% | 3 | India | 7.03% | 3 | |
| Bahamas | 0.93% | 4 | Timor-Leste | 7.20% | 4 | |
| Brunei | 1.25% | 5 | Comoros | 8.84% | 5 | |
| Haiti | 1.35% | 6 | Burkina Faso | 9.02% | 6 | |
| Kuwait | 1.69% | 7 | Brazil | 9.24% | 7 | |
| Italy | 1.75% | 8 | Argentina | 9.35% | 8 | |

| | | | | | |
|-------------|-------|----|---------------|--------|----|
| Switzerland | 1.83% | 9 | Ethiopia | 9.36% | 9 |
| Germany | 1.87% | 10 | United States | 9.66% | 10 |
| Russia | 1.87% | 11 | Burundi | 9.68% | 11 |
| El Salvador | 1.96% | 12 | Afghanistan | 9.74% | 12 |
| N Macedonia | 1.99% | 13 | Bangladesh | 9.84% | 13 |
| Eswatini | 2.02% | 14 | Eritrea | 10.24% | 14 |
| Puerto Rico | 2.06% | 15 | Sudan | 10.91% | 15 |
| Belgium | 2.09% | 16 | Rwanda | 11.29% | 16 |
| Lesotho | 2.12% | 17 | Nepal | 11.51% | 17 |
| France | 2.13% | 18 | Japan | 12.27% | 18 |
| Argentina | 2.23% | 19 | Pakistan | 12.34% | 19 |
| Portugal | 2.24% | 20 | Guinea-Bissau | 12.77% | 20 |

Bottom 20 countries wrt to growth rates compared to bottom 20 exporters

COMPARISON OF PERCENTILE RANKS WITH RESPECT TO GROWTH AND EXPORTS

We start with a list of the top 20 and the bottom 20 countries with respect to growth rates. For each of the 40 countries, we also list their percentile ranks in terms of their export/GDP ratio. This list clearly shows the lack of association between growth rates and exports. There are a few countries in the top 20 which have excellent performances in terms of export to GDP ratio – for example, Qatar, Macao, Singapore, Djibouti, Malaysia, and Angola are above 80th percentile in Exp/GDP ratio. However, many of the countries in the top 20 have very poor export performances. Similarly, among the bottom 20, the worst performers in terms of growth, we also have six countries which are above 80th percentile in terms of Exp/GDP ratio. Having an excellent export performance did not prevent them from ranking in the bottom 20 in terms of growth.

Similarly, looking at the top and bottom 20 countries in terms of exports to GDP ratio and evaluating these countries in terms of their growth performance does not reveal any relationships. The best performance on export/GDP ratio can correspond to any kind of growth performance – good, neutral and bad. Similarly, the worst export/GDP ratios can have any kind

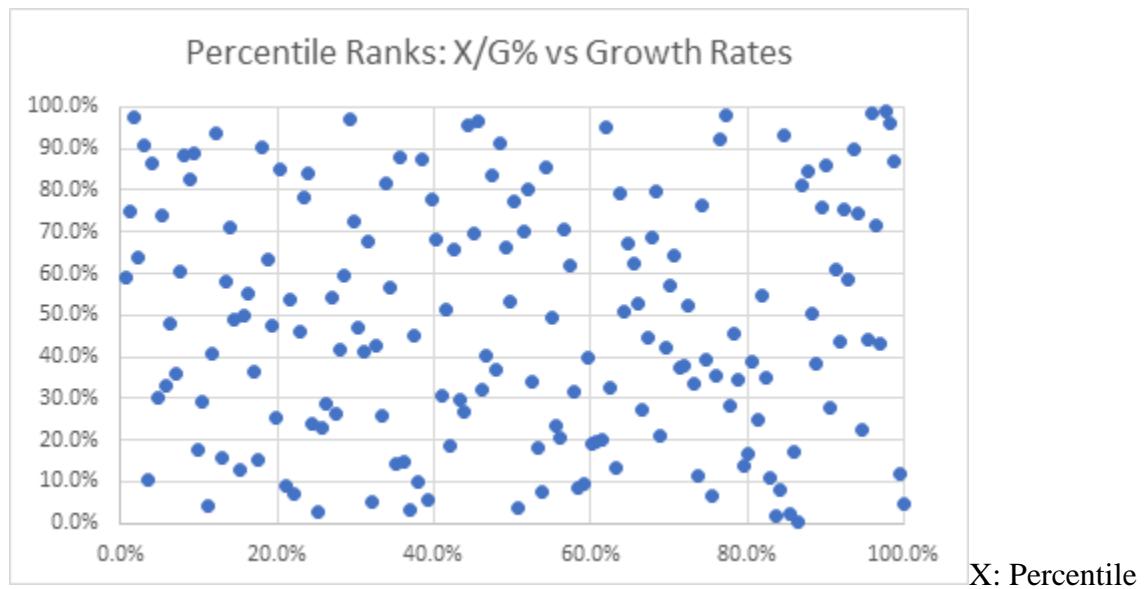
of growth performance. If there is a relationship between growth rates and exports, it is not very strong.

| Top 20 | Growth h | %-rank | X2G | Bottom 20 | Growth h | %-rank | X2G |
|------------------|----------|--------|------|-----------------|----------|--------|-------|
| Country | Gr Rnk | rnk | | Country | Gr Rnk | rnk | |
| Ethiopia | 100.0% | | 4.5% | Jamaica | 0.6% | | 59.2% |
| China | 99.4% | % | 11.9 | Palau | 1.2% | % | 75.1% |
| Qatar | 98.8% | % | 87.1 | S. Maarten | 1.8% | % | 97.5% |
| Macao | 98.2% | % | 96.0 | Bahamas | 2.3% | % | 63.7% |
| Singapore | 97.7% | % | 99.0 | Brunei | 2.9% | % | 91.0% |
| Korea, Rep. | 97.1% | % | 43.3 | Haiti | 3.5% | % | 10.4% |
| Tajikistan | 96.5% | % | 71.6 | Kuwait | 4.1% | % | 86.6% |
| Djibouti | 95.9% | % | 98.5 | Italy | 4.7% | % | 30.3% |
| Lao PDR | 95.3% | % | 44.3 | Switzerlan d | 5.3% | % | 74.1% |
| Kiribati | 94.7% | % | 22.4 | Germany | 5.8% | % | 32.8% |
| Azerbaijan | 94.2% | % | 74.6 | Russia | 6.4% | % | 47.8% |
| Malaysia | 93.6% | % | 90.0 | El Salvador | 7.0% | % | 35.8% |
| Cote d'Ivoire | 93.0% | % | 58.7 | Macedoni a | 7.6% | % | 60.7% |

| | | | | | | | |
|------------|-------|---|------|-------------|-------|---|------|
| Cambodia | 92.4% | % | 75.6 | Eswatini | 8.2% | % | 88.6 |
| Armenia | 91.8% | % | 43.8 | Puerto Rico | 8.8% | % | 82.6 |
| Greenland | 91.2% | % | 61.2 | Belgium | 9.4% | % | 89.1 |
| Ghana | 90.6% | % | 27.9 | Lesotho | 9.9% | % | 17.4 |
| Angola | 90.1% | % | 86.1 | France | 10.5% | % | 29.4 |
| Botswana | 89.5% | % | 76.1 | Argentina | 11.1% | | 4.0% |
| Mozambique | 88.9% | % | 38.3 | Portugal | 11.7% | % | 40.8 |

Growth and Export %-ranks for top and bottom 20 countries wrt growth rates

Perhaps the relationship is not strong at the bottom and top, but works well in the middle? To assess this, we look at a scatter-plot of percentile ranks in both factors – exports and growth rates. This is plotted below:



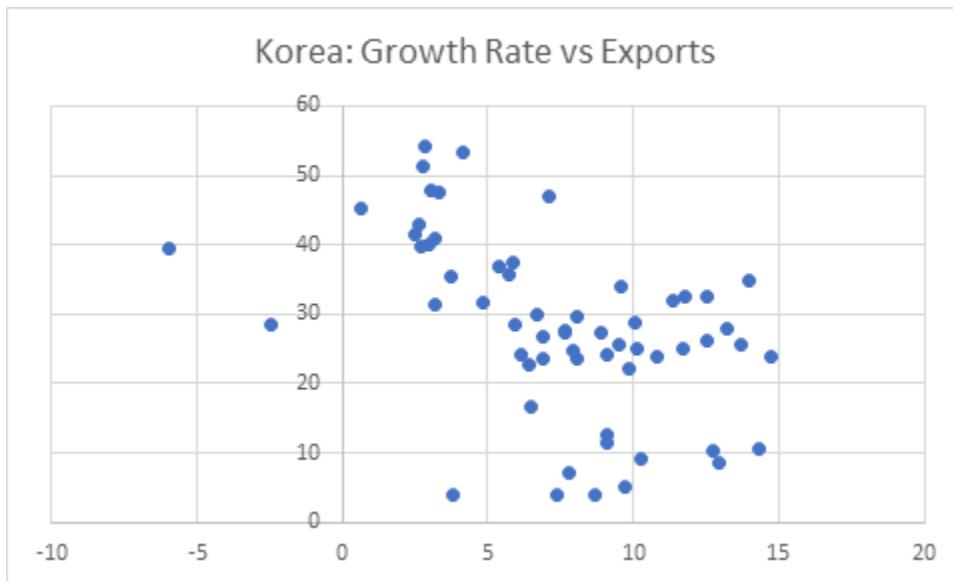
Growth Rank Y:Percentile Export Rank

This plot shows a complete lack of relationship between exports to GDP ratio and the GDP growth rate.

THE CASE OF KOREA:

There is another way to think about the relationship between exports and growth rates. Instead of looking at export performance in comparison to other countries of the world, we could just confine comparisons to domestic performance. Is it the case that, within a given country, periods of high growth correspond to periods of high exports? This question is different from the one we have been examining before, where high growth and high exports are defined relative to global performance. Here performance comparisons are solely internal, and high growth just means higher growth for that country relative to its own growth performance.

To examine this question, we take the case of one country, namely Korea. For this country, we have a nearly complete data set. Korea was also a star performer with respect to exports over a certain period of time, although not throughout the period for which the data is available. This makes it a good test case. Below is an X-Y plot of the growth rates and the export to GDP ratio for Korea. This plot shows no relationship between exports and growth rates. At the lowest growth rates, export percentages vary from the smallest to the highest. The range of variation is reduced at the highest growth rates, but the export percentages are all in middle ranges. That is, highest growth rates correspond to periods with the export ratios are in the intermediate range, neither too high, nor too low. Again, this confirms the basic lesson that there is no relationship between exports and growth.



CONCLUSIONS:

What do we learn from this data analysis? The data shows that there is no association between export to GDP ratios and growth rates. What causal conclusions follow from this lack of association? It is clear that there is no strong relationship between the two variables. The hypothesis of export-led growth has many devoted supporters, who could find a thousand ways to dispute the above analysis. For example, we could begin by disputing the WDI data. Why does WDI have zero data on growth rates for more than 40 countries? There are many missing growth rates even for countries for which WDI gives a complete series for the GDP! There are

many different GDP series, and one could make arguments for using any one of them. Similarly, there are many different export series, and switching series might lead to different conclusions. Instead of pursuing this line of analysis, we turn to discuss the META-question: “Why are we discussing export-led growth?”. To understand this, we quote from a Brookings Review of a book entitled “The Key to Asian Miracle”:

“Eight countries in East Asia—Japan, South Korea, Taiwan, Hong Kong, Singapore, Thailand, Malaysia, and Indonesia—have become known as the “East Asian miracle” because of their economies’ dramatic growth. In these eight countries real per capita GDP rose twice as fast as in any other regional grouping between 1965 and 1990. Even more impressive is their simultaneous significant reduction in poverty and income inequality. Their success is frequently attributed to economic policies, but the authors of this book argue that those economic policies would not have worked unless the leaders of the countries made them credible to their business communities and citizens.”

These East Asian countries made dramatic progress in growth and managed to transition from agricultural economies to industrial economies. The question of how they managed this successfully, when nearly all over the globe, other countries trying to achieve this transformation, failed to do so? Some of the observers noted that all of these countries achieved very high exports, and hypothesized that this was the key to the rapid growth. The data shows that this hypothesis is far too simple as an explanation. There is no simple formula for rapid growth, and there are many examples of rapid growth without exports. In fact, growth depends on increasing domestic productive capacity. When this productive capacity increases, it can be channeled to exports, or simply used to increase domestic consumption. The World Bank data strongly rejects the export-led growth hypothesis, as per our data analysis given in this lecture.

9C: The Association between Income and Consumption

In previous parts ([7A](#) & [7B](#)) of this lecture, we saw that exports/gdp ratio and gdp growth rates had no relationship to each other. In this part, we examine the opposite case. According to economic theory, the two series of GDP (Y) and CONS (C) are closely related to each other. We will examine the WDI data to assess the strength of the relationship between the two series.

Preliminaries: The Consumption Function

The Consumption Function is a Fundamental Building Block of Keynesian Economics. Consumption is a function of the Income. This is justified as a BEHAVIORAL relationship. I earn income, and I consume some proportion of it, and save the rest. Because of this theoretical basis, the relevant variables to study are the GDP/capita and CONS/capita. The GDP is the total income of the country, while the per-capita reduces this to income in hands of individuals. Similarly consumption per capita reflects average individual consumption. This reduction is not very good because the national income is NOT equally distributed among all people. It would be much better to study these variables separately in quintiles. That is, divide income earners into five classes – poorest, below average, average, above average, richest – and then study these classes separately. Some data is available on this basis, but there is very little, so it would be hard to do a systematic study. Therefore we ignore this problem and proceed using GNP per capita while recognizing that this is not adequate, because the relationship between consumption and income CHANGES within these five different income classes.

Another issue of great importance is the CAUSAL DIRECTION of the relationship. Keynesian theory says that $GDP \Rightarrow \text{CONS}$, and NOT the other way. However, DATA can ONLY tell us about the association between the two variables, and not the direction of the causal arrow. This is an important issue and will be discussed in greater detail later.

Relevant Variables from WDI Data Base

Upon examination of the WDI data base, we find two variables, listed below, which appear directly relevant to our research. These are listed below:

NY.GNP.PCAP.KD Economic Policy & Debt: National accounts: US\$ at constant 2010 prices: Aggregate indicators GNI per capita (constant 2010 US\$) *GNI per capita is gross national income divided by midyear population. GNI (formerly GNP) is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output plus net receipts of primary income (compensation of employees and property income) from abroad. Data are in constant 2010 U.S. dollars*

NE.CON.PRVT.PC.KD Economic Policy & Debt: National accounts: US\$ at constant 2010 prices: Expenditure on GDP Household final consumption expenditure per capita (constant 2010 US\$) *Household final consumption expenditure per capita (private consumption per capita) is calculated using private consumption in constant 2010 prices and World Bank population estimates. Household final consumption expenditure is the market value of all goods and services, including durable products (such as cars, washing machines, and home computers), purchased by households. It excludes purchases of dwellings but includes imputed rent for owner-occupied dwellings. It also includes payments and fees to governments to obtain permits and licenses. Here, household consumption expenditure includes the expenditures of nonprofit institutions serving households, even when reported separately by the country. Data are in constant 2010 U.S. dollars.*

Real Statistics involves knowing the what the numbers mean in the real world.

Accordingly, we provide a brief explanation of what these series measure, before proceeding to check how strongly they are associated with each other.

Creating Comparability Across Countries and Across Time

Briefly, GDP (Gross Domestic Product) is the market value of final goods and services traded in market place. Intermediate goods, produced for use in producing other goods do not count. Technically, we must count the money traded at of all market transactions where money is exchanged between consumers and producers. Practically, this is impossible to do, and various kinds of approximations are made to arrive at an estimate of this. It useful to recognize that there are two approaches to the measurement of GDP. The Expenditure Approach attempts to measure the amount of money consumers spend on purchases of goods and services. The Income Approach attempts to measure the amount of money businesses earn by the sale of goods and services. Of course, these two numbers should be exactly the same, but data may be available on one side and not on the other.

The end product of any attempt to measure GDP (now called Gross National Income or GNI in WDI accounts) will be a number in LCU (local currency units) of the total amount of sales of goods and services in the marketplace. Thus, Chinese GDP will be in Yuan, while the Swedish GDP will be Kroners. Obviously, these numbers would not be comparable. In order to be able to compare GDP across nations, we need to put these into common units. For this purpose, we can

convert from LCU to USD US Dollars. However, this is by no means a simple process. The exchange rate varied on a daily basis, and sometimes fluctuates a great deal from the beginning to the end of the year. If the exchange rate of US \$1 for PKR varies between 100 and 150 in a given year, the GDP translated into USD could change by 1/3, which is a huge amount. The World Bank uses a complex method, called the Atlas Method, to carry out the conversion in a way that creates a stable series of measures, which takes the changing exchange rates into account properly. It is worth noting that all such conversions are based on guesses and approximations, and may not work very well, especially in extreme situations.

Converting LCU to USD creates comparability across countries, but not across time, because the value of the dollar fluctuates over time. We can create a suitable price index which measures the price of dollars across time. Setting the value of the dollar to 100 in 2010 and then dividing the current dollars by the price index creates a constant 2010 dollar valuation of the GDP which is comparable across time and across countries. The IDEA of these transformations is to measuring the QUANTITY of goods being produced – not the VALUE = $P \times Q$. However, this IDEAL can NEVER be actually achieved, for a wide variety of reasons. The point here is to recognize that there are MANY problems with the numbers given in WDI, and if we want to pursue any specific issue, we may well be able to dispute this numbers and create better alternatives focused on that particular issue. Currently, we are pursuing the general theme of relationship between consumption and income, and we can ignore these problems for this purpose.

Some Additional Real-World Issues

These issues are peripheral to our main concern at present, but worth clarifying nonetheless. Since the consumption function comes from behavior of families (individuals), aggregated over the nation, the GDP per capita measure seems more appropriate than total GDP. However, if we were investigating wealth of nations, in terms of global power, the gross GNP would be more suitable. For investigating welfare, it is more suitable to look at the family incomes. As already remarked, GDP per capita is a very poor measure of this, because it measures how much families would have if the GNI was distributed equally among all families, and ignores inequality.

Another adjustment of great importance if we want to compare household welfare across nations is the issue of purchasing power parity. Although we have equalized incomes by measuring them in constant 2010 dollars, the purchasing power of the dollar varies across countries. The same household goods are very expensive in some countries and very cheap in others. Adjustments for this can be made, and go under the name of PPP or Purchasing Power Parity. However, this data set is limited, so we will not use it in our current investigation.

Private Consumption per Capita: NE.CON.PRVT.PC.KD

This is the total value of goods and services sold to the private sector, divided by the population, to put it in per capita terms. There are many complexities and ambiguities in both the definition and the data collection for this variable. For example, durable goods like houses are not consumed in one-year. Instead, one should compute value of services obtained for the good for one year, and so on. Much more important is the fact that there are many very important “social” transactions in a society which do not go through the marketplace. Women provide childcare and housekeeping services which are mostly unpaid. When women work get a job, they earn income, and also they hire housekeepers and childcare, so there is double-increase in GNP. But it is very likely that social welfare goes down, because the value of social services provided is far greater

than the market can pay for.

But perhaps the BIGGEST problem is that a PER CAPITA hides INEQUALITY. That is, instead of measuring the dramatically different incomes and consumption levels of different segments of society, it aggregates everyone into one collective, and takes the average over all people in society. We have already discussed how dividing people into five different groups (quintiles) by income would help solve the problem. Another method would be to divide the society in social classes, like capitalists and laborers, or along other lines, and study incomes and consumptions in different groups separately.

The Politics of Data

One of the GOALS of CAPITALIST economic theory is to HIDE class conflict, power, and inequality. It is VERY EXPENSIVE to gather data on a global basis. The data that is gathered reflects the interests of the powerful. The data that is NOT gathered also reflects political interests. Angus Deaton, recent Nobel Laureate, has expressed this point clearly in many of his essays. This is why we have statistics on GDP per capita, but very little data on poverty, inequality, and on distribution of power. In this connection, it is important to note that our analyses of GDP per capita distracts attention from inequality and poverty. This is helpful in maintaining existing power structures.

In general, the job of the academia and research is to support existing power structures and to suppress revolt and dangerous ideas, by branding them as heretical and unworthy of spending time on. For example, Nobel Laureate Lucas, a champion of Chicago School free market economics, asserts that “Of the tendencies that are harmful to sound economics, the most seductive, and in my opinion the most poisonous, is to focus on questions of distribution. The potential for improving the lives of poor people by finding different ways of distributing current production is nothing compared to the apparently limitless potential of increasing production.” Lucas tells us to focus on growth and forget about inequality, basically advocating the trickle-down theory of poverty reduction. In contrast, a World Bank research paper by Cordoba and Verdier on “Lucas vs Lucas” uses Lucas’s own welfare framework to assess this claim, and finds that focus on inequality would yield far greater benefits to the poor, compared to growth.

Data Analysis: Cross-Section, Time Series, and Trends

From these deep and difficult issues, we come back to a simple question. How can we check if GDP per capita (NY.GNP.PCAP.KD) and CONS per capita (NE.CON.PRVT.PC.KD) are associated? In lecture 7B, we checked Exports/GDP and GDP growth rates for association. Can we use the SAME method we used for Exports and Growth? The answer is “NO!” To understand “Why Not?” we must discuss different types of data. We give a brief description of four types of data sets.

The simplest kind is a Cross-Section Data Set. One example this is data on a collection of countries at ONE point in time. The second type is a Time Series. One example is data on one variable taken in a sequence of points of time. There are two important types of combinations. One of them is a collection of cross sections at different points of time, in sequence. Here the individuals in the cross-section may not be identified, and the cross section at one point may be of different individuals (or countries) than at a later point of time. The best type of data set is a PANEL data set, where we have a cross-section of data which is repeated across time, with identified individuals. This allows us to get a time series for each individual, and also get a cross-

section of the all the individuals at any given point in time. The WDI data set is a Panel Data Set. We have the cross section of all countries, and we have time series for each country. This is best because it gives us the maximum amount of information.

Cross-Section data has a very important property. The Ordering of the data does not matter; the sequence in which the data is presented does not carry information. For example, we can arrange the countries in alphabetical order, or in random order, or sort the data set according any variable. The data set will contain the same information. There can be no trends in the data sets, because we can arrange data as we like.

Why Median Does Not Properly Rank Time Series With Trends

In contrast to this, in a Time Series, sequencing is of great importance. The order of the data carries very important information about long run tendencies of the variables under study. IN particular, GNP and CONS have increasing trends. All countries are making strenuous efforts to achieve growth in GNP, and populations also growing, requiring greater levels of production. Thus, with few exceptions, the GDP numbers as well as the consumption numbers grow systematically over time. In contrast with this, the Exp/GDP ratio has no systematic trend across time. For each country, we can see it rise and fall without any systematic patters. The same is true for the rate of growth of GDP. This does not show any systematic trends across time, and rises and falls for nearly all the countries. To see why this matters, consider what would happen if we apply the same method for judging association that we did in Lecture 7B.

In 7B, we used the Median Exp/GDP ratio to classify countries according to whether they were big on exports or not. What would happen if we used Median (GDP) to classify countries into high GDP and low GDP? We have approximately 60 points of data from 1960 to 2019. If the data is increasing, then the midpoint will occur in the middle of the data set, so the Median(GDP) will be approximately GDP(1989). Similarly, median(CONS) will be CONS(1989). So the median will be the value in the middle year, and will not represent the 60 years of data in any way. So how can we classify the data into high and low GDP countries, when GDP changes over time, and the ranks of different countries are also changing across time? Actually, this is quite a difficult question. Instead of answering it, we will avoid it, and use a different method.

An year-by-year analysis

Instead of judging high and low GDP countries across time, let us just fix one point in time. We start with the year 2018, as this is the most recent year. Now it is easy to rank countries according to GDP, since this is now a cross-section of countries, and each country has just ONE GDP number. For a cross-section with no trends, we can follow the methodology already developed in Lecture 7B, and pick out the top 20 countries by GNP and also, separately, by CONS. Doing so, we get the following table:

Top 20 w.r.t GNP & CONS in the Year 2018

| | Top 20 | | %-ranks | Top 20 | |
|---------|---------------|--------|---------|---------------|--------|
| Country | GNP | CONS | Country | GNP | CONS |
| NOR | 100.0% | 100.0% | NOR | 100.0% | 100.0% |

Author Last Name/Book Title

| | | | | | |
|-----|-------|-------|-----|-------|-------|
| LUX | 99.3% | 98.6% | USA | 95.7% | 99.3% |
| DNK | 98.6% | 96.4% | LUX | 99.3% | 98.6% |
| IRL | 97.8% | 90.6% | AUS | 94.9% | 97.8% |
| SWE | 97.1% | 94.9% | CAN | 92.8% | 97.1% |
| NLD | 96.4% | 89.1% | DNK | 98.6% | 96.4% |
| USA | 95.7% | 99.3% | GBR | 87.7% | 95.7% |
| AUS | 94.9% | 97.8% | SWE | 97.1% | 94.9% |
| SGP | 94.2% | 86.2% | HKG | 86.2% | 94.2% |
| MAC | 93.5% | 80.4% | JPN | 92.0% | 93.5% |
| CAN | 92.8% | 97.1% | FIN | 90.6% | 92.8% |
| JPN | 92.0% | 93.5% | AUT | 91.3% | 92.0% |
| AUT | 91.3% | 92.0% | DEU | 89.9% | 91.3% |
| FIN | 90.6% | 92.8% | IRL | 97.8% | 90.6% |
| DEU | 89.9% | 91.3% | BEL | 89.1% | 89.9% |
| BEL | 89.1% | 89.9% | NLD | 96.4% | 89.1% |
| FRA | 88.4% | 87.7% | NZL | 84.8% | 88.4% |
| GBR | 87.7% | 95.7% | FRA | 88.4% | 87.7% |
| ARE | 87.0% | 79.7% | ITA | 84.1% | 87.0% |
| HKG | 86.2% | 94.2% | SGP | 94.2% | 86.2% |

Top 20 according to GNP/cap and CONS/cap, with %-ranks

Unlike the case in lecture 7C, where there was no overlap between the top 20 wrt exports & growth rate, here 18 out of 20 countries appear in both top 20 lists. The two countries which are different are also close to being in the top 20. In particular MAC = Macau, ARE = United Arab Emirates are in the top 20 GNP only, while ITA = Italy, NZL= New Zealand top 20 CONS only. But the percentile ranks are within 10% of each other. This means the high rank in GDP ensures high rank in CONS and vice-versa. Thus the two variables are closely associated. To check this further, for all countries instead of the top 20, we look at the difference in %-rank between a GDP ranking and a CONS ranking. Examining the list, we find that there are four countries which have a difference of greater than 10%. These are Republic of Congo, Brunei,

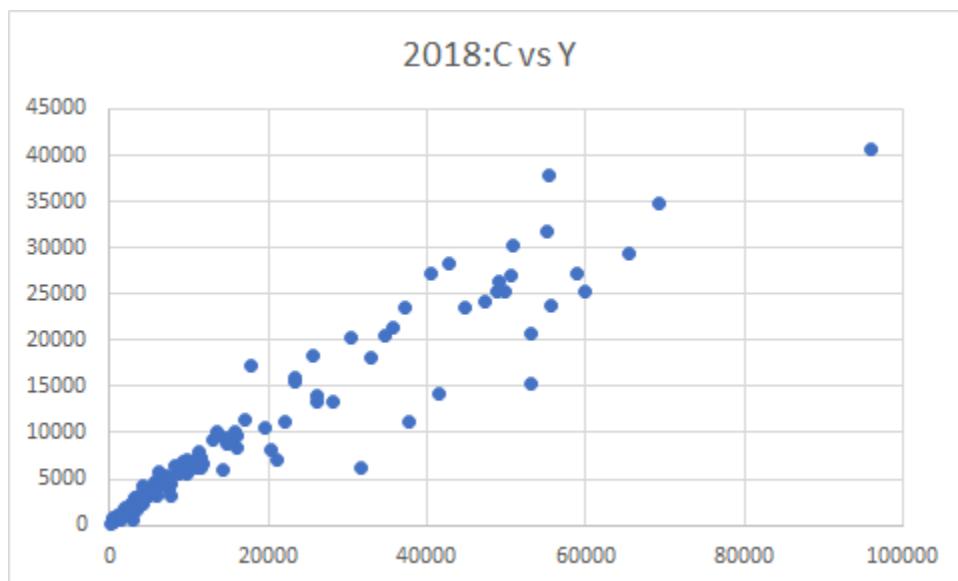
China, and Macao. All of these countries have high %-ranks on GDP, but much lower %-ranks on CONS. Brunei and Congo are major oil exporters, which creates high GDP, but the wealth is not used to provide consumption to the public, which accounts for the disparity. Similarly, Macao, is a special administrative territory within China, which enjoys extremely high levels of GDP due to foreign trade, but also has high inequality. China is a special case. It has a one-party rule which has created massive growth, while maintaining low consumption profiles enabling high savings needed for growth.

The above discussion illustrate a general methodological principle of real statistics. Exceptional differences in %-ranks point to a real world reason, but the data do not contain the reason. To find out WHY, we must go beyond the data to the real world causes of this exception.

A Graphical Analysis of 2018 GDP and CONS

We have used percentile ranks to demonstrate a strong association between GDP and CONS. However, sophisticated analysis is not necessary. A simple graph of the data – directly – without any calculations – provides strong evidence of association.

Graph of Private Consumption per capita, plotted against GDP per capita, for 132 countries in WDI



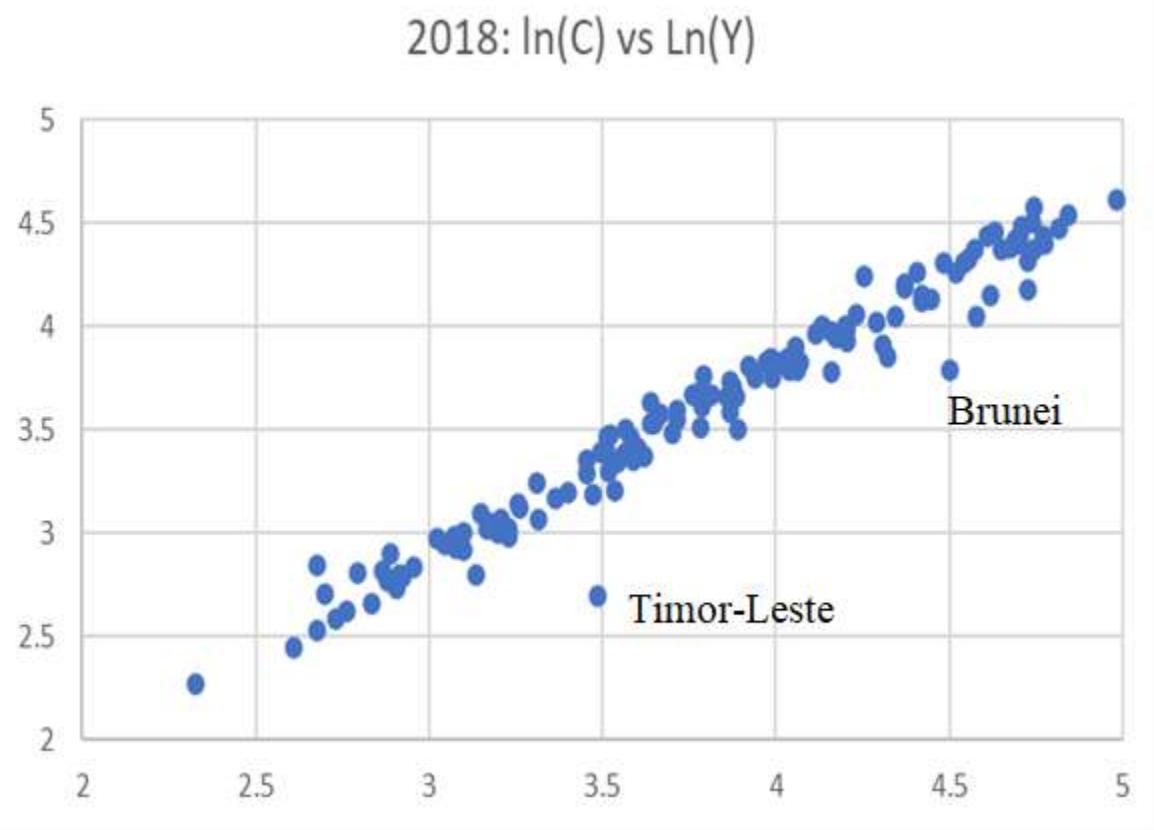
This shows clearly that as GDP rises, CONS also rises. Note the contrast with the graph in previous Lecture 7C of Exp/GDP ratio and GDP growth rates. In the previous chart, the dots were scattered evenly all over the place, showing lack of relationship. The graph is very tightly clustered in the early portion, and becomes looser in the later part, but the strong association is clear.

Benefits of Log Transformation

Generally speaking, income distributions tend to have right outliers. That is, a small percentage of incomes are extremely high, while a large majority are very low. This type of distribution is called a Pareto Distribution. From the graphical point of view, the high Incomes increase the scale of the graph, and make the patterns in the small incomes hard to see. Also, the variability of consumption as a proportion of income is higher at high income levels, which makes the

relationship appear to be somewhat erratic at higher income levels. As we shall see, these problems can be solved by making a log transformation in the data.

There is also a theoretical reason why a log transformation may be useful for this data set. Economic theories suggest that Consumption and Income are roughly proportions, so that $C/Y=K$. This implies that $\log C = \log Y + \log K$. Proportionality translates to a linear relationship of slope 1 between the logs of C and Y. Linear relationships are easier to spot visually. In addition, the log transform removes scale distortions discussed earlier. The plot of Log (C) versus Log(Y) displays these benefits.



Strong association between the two variables is evident from graph. The log scale improves visibility of lower points. The nearly linear relationship between the two variables is also clearly displayed in the graph. It can be seen that the SLOPE of line is 1; that is, the line shows that C is proportional to Y across countries in this sample. The high variability of consumptions at high incomes has also been removed by the log transformations. The outliers in the graph are different from the ones that appear in the earlier graph, without the log transformation. The four countries which appear as the biggest outliers with respect to ranks do not show up so strongly in the log transformed original data. Instead, to the eye, Timor-Leste is the biggest outlier. With $\log(\text{GDP})= 3.5$, $\log(\text{CONS})=2.7$ is much lower the average value of 3.5. This shows up as an outlier on the graph. One can investigate the special aspects of Timor-Leste economy to explain why the GDP is high, but CONS is not high in the same proportion as for other countries.

What about Other Years?

So far, we have only analyzed one year of data, namely that for 2018. What happens in other years? Do we get similar results? It turns out that the association between GDP and CONS is very strong throughout the period for which we have data. There are many ways to confirm and verify this claim. The slowest and most painful would be to go through this analysis for each year separately. In a class, this would actually be useful to do by assigning every student a different year to analyze, using the same methods as discussed above. A fast method would be to create percentile ranks for each country for each year, and look at the differences between the GNP percentile rank and CONS percentile rank over the entire data set. After removing a small number of outliers, corresponding to exceptional circumstances, the biggest difference in percentile ranks is less than 10% over nearly the entire data set. This shows the strong correlation between the two variables over the entire time period for which we have data.

The Question of Causality

The more important question is the causal relationship. WHY is there such a strong relationship between GNP and CONS? Is it because GNP \Rightarrow CONS, or is it the other way around? Or, is some other, more complex, causal mechanism at work, which creates a strong relationship between the two. In answering this question, the data available cannot help us. Rather, we need to know more about the causal structures of the world we live in, and how these are reflected in the data we have. The data can only show us the symmetric associations, while causality is usually one directional and asymmetric. We will discuss the causality question in greater detail in later lectures. For the moment, we are only concerned with a surface analysis of what the data shows about the symmetric associations between pairs of variables.

Association of C & Y within one country

We have looked at the relationship between consumption and income on a global level, and deduced association from that fact that for both variables, different countries have roughly the same rank among all the countries for which we have data. This is very different from the standard approach to this topic. In general, econometricians study the relationship between GNP and CONS for one country, without reference to global data. In fact, there are very difficult problems which arise if we try to establish that there is an association between GDP and CONS for a single country. At the moment, there does not exist any fully satisfactory resolution of this problem within conventional econometrics, even though it has been studied intensively for many decades. The problem can be stated simply. Any method used to show that GDP (TURKEY) and CONS (TURKEY) are associated, also ends up showing the GDP(Nepal) and CONS(Uruguay) are associated. The second is an example of what is called “spurious correlation” or “spurious regression” or “nonsense regression”. That is, statistical analysis shows a strong relationship, whereas there is no relationship between the two variables in the real world. This is because both variables have increasing trends in time.

Why Time Trends Cause Difficulty in Assessing Association

We first explain how our method of judging association will lead to false and misleading results in this case. To be specific, consider the case of Australia. Both Consumption & Income have increasing trends over time. Which are the years for which Australian GDP would be in the top 20. Because of the increasing trend, these would obviously be the LAST 20 years in the data set,

from 2000 to 2019. Similarly for CONS, the top 20 years would also be the last 20 years, since it also has an increasing trend. The bottom 20 years for both variables would be the first 20 years, from 1960 to 1979. So there would be a near perfect match between the top 20 and the bottom 20 when we look at the 60 years of data. But this match only reflects the common trend in the two variables and does not tell us anything about the association. To see this clearly, consider some series which is totally unrelated to Australian GDP but also has a time trend. For instance, consider the GDP series for Peru, which (like nearly all countries of the world) also has an increasing trend. We can ask if the two series are associated. Our method will give us the result the “YES” these two series are associated, because they are both very high in the same time periods (after 2000) and they are both very low in the same time periods (before 1980). But in fact these series have no relationship to each other. The problem lies in our use of the WRONG method for judging association. This method leads to the result that any two series with increasing time trends will be associated. This is obviously a wrong conclusion.

In fact, standard methods for checking association for Time Series also work very poorly. It turns out that we need DIFFERENT methods, which are a bit complex, to assess association for two Time Series variable which both have trends. This question will be considered in greater detail later.

Concluding Remarks

As we have seen, checking associations between two variables is NOT a mechanical task. How to check associations depends on the nature of the data. The methodology we have developed in lectures 7B and 7C is suitable for cross-sections. It can also work for time series if they DO NOT HAVE TRENDS. But It DOES NOT WORK if the time series have trends. Conventional Statistical methods also fail to assess correlation/association correctly in times series with trends. Finally, causality is not DIRECTLY visible in data, but we do get hints. For example, in the case of Export-Led growth, the data strongly rejects the hypothesis, showing us clearly that there is no relationship between Export/GDP ratio and the growth rates. The lack of association allows us to (tentatively) reject the hypothesis of causality from exports to growth. This rejection is tentative because a REFINEMENT of the hypothesis might work. We might only consider “industrial exports” as being relevant, for example. This is how science works. We start with an idea, and check against the data. If the data confirms it, then our belief in the hypothesis is strengthened. If the data rejects, we go back to the drawing board.

9D: Testing a Coin for Fairness

A fair coin is one for which the probability of both outcomes – H (heads) or T (tails) – is 50%. That is, both outcomes are equally likely. Given a coin, how do we determine whether or not it is a fair coin? Another way to think about this question is the following. Suppose we observe a deviation from fairness – either too many heads or too many tails. How do we decide whether this deviation is due to chance, or whether it reflects a bias in the coin?

Philosophy of Probability

Before proceeding, we pause to reflect on the concept of a fair coin. What does it mean to say that the chance of the outcome H is 50%? Perhaps surprisingly, this is a deep and difficult question, which has created bitter disputes which continue to this day. The reason for the

problem is that the probability is unobservable. When we say that the probability is one half, this means that even though the coin came up heads, *it could have come up tails*. This is a counterfactual – something which did not happen, but could have happened. By definition, this cannot be observed. According to Logical Positivist philosophy, dominant at the time probability concepts were being formulated, scientific knowledge cannot be built on unobservables. Accordingly, a lot of effort was made to avoid counterfactuals, and to define probabilities in terms of observables. All of these efforts failed, but the results of these failures are bad theories of probabilities which continue to confuse minds and dominate the textbooks.

Although the philosophical foundations are very important, and worth discussing in greater depth, we will bypass them for now and take an intuitive approach. We will assume everyone knows, intuitively, the meaning of equally likely events, and also the counterfactual that the coin could have come up heads, even though it came out to be tails. Given this simple and intuitive definition of probability, we can proceed as follows.

Pragmatic Solutions:

A simple way to assess whether or not the coin is fair is to just flip it a few times and note the outcomes. For a fair coin, all sequences of outcomes are equally likely. For example, in two flips, the following four sequences can occur: TT, TH, HT, HH. All four have equal probability of 25%. This means that if we flip a coin twice and see two heads, we cannot conclude that the coin is biased towards heads. When the number of flips is small, chance variations can create large numbers of Heads or Tails. So to assess whether or not a coin is fair, we need to flip it a large number of times. There is a theorem called The Law of Large Numbers which says that the proportion of heads will be close to 50% when the number of flips is large, if the coin is fair. We cannot use this theorem to create a perfect test for fairness because both “large number of flips” and “close to 50%” are imprecise terms. But the theorem does give us an idea that we should test using a large number of flips. To understand the problem better, we will consider testing the coin for fairness using 10 flips of the coin.

Basic Counts & Probabilities:

We want to study what happens when we flip a fair coin ten times. In particular, we would like to learn about the NORMAL and usual behavior of such sequences of coin and also learn about the ABNORMAL and unusual behavior. This allows us to test for fairness. If the coin behaves in a way that is abnormal for fair coins, we will guess that the coin is probably not fair – that explains the deviations from fairness that we observe in the behavior. To learn about what is normal and what is not, we need to study the probability of outcomes for different numbers of heads in the sequence. These probabilities are called binomial probabilities. It is not our goal here to study these probabilities, and how they are calculated, in any detail. We will provide a brief introduction to the topic, and use the computer to calculate the probabilities.

For a sequence of ten flips, each flip has two possible equally likely outcomes: H or T. In ten flips, we have a total of $2 \times 2 \times \dots \times 2$ or $2^{10} = 1024$ possibilities. The key to the probability calculations is the understanding that all of these 1024 outcomes are equally likely. That is, each one has probability 1/1024. Any arbitrary sequence like HTTTHHTHTT has probability 1/1024

because each one of the two H,T has equal probability in all of the ten places in the sequence. In order to calculate the probability that the number of heads is K, where K is an integer between 0 and 10, we just need to COUNT the number of sequences which have K heads. We show how to do this in some simple cases.

The simplest case is when you have 10 heads. There is only one such sequence: HHHHH HHHHH. This sequence has probability 1/1024. The counterpart to this is when you have 0 heads. Again there is only one such sequence: TTTTT TTTTT. This sequence also has probability 1/1024. There is a fundamental symmetry between heads and tails. The probability of K Heads is the same as the probability of K tails, but K tails corresponds to 10-K Heads. So the probability of K heads is the same as the probability of 10-K heads. Note that both of these events are highly unusual and have very low probabilities. We do not expect a fair coin to come out all heads or all tails, since this would display heavy bias. The probability calculation shows that these events have probability less than 1 in 1000.

Next we consider calculating the probability of exactly 9 heads. How many sequences are there with 9 heads? The sequences must have exactly one T in them. This one T can be in any one of 10 positions: THHHH HHHHH, or HTHHH HHHHH, or HHTHH HHHHH, and so on. So there are 10 sequences with exactly 9 heads. This means that the probability of exactly 9 heads is 10/1024, which is around 1%. So this is still a very low probability, and a very unusual event for a fair coin. If we see 9 heads and 1 tail, or 9 tails and 1 head, we are likely to reject the idea that the coin is fair.

Next consider the case of 8 heads. How can we count the number of sequences which have TWO tails and 8 heads in them? This can be done in the following way. The FIRST tail can be put in 10 different places. The SECOND tail can be put into the remaining 9 places. So there are a total of $10 \times 9 = 90$ ways of putting in a FIRST tail and then a SECOND tail into a sequence of 10 positions. We can use a shorthand notation to indicate this. {1,2} means that we put the first tail in the first position and the second tail in the second position, getting the following sequence: TTHHH HHHHH. Similarly, {3,9} means: HHTHH HHHTH. The 3 and 9 indicate the position of the first and second tail. Any pair of numbers like {6,3} corresponds to positions for the first and second Tail. The first number can range from 1 to 10. The second number must be different from the first, and so it has 9 possibilities. This gives us $90 = 10 \times 9$ possibilities. However, this procedure involves double counting. The sequence {2,3} is the same as {3,2} – both put tails in the second and third position: HTTHH HHHHH. The reason for this double counting is that we consider the FIRST tail as being different from the SECOND tail. The {2,3} sequence has H T1 T2 H H H while the {3,2} sequence has H T2 T1 H H H. The order in which the tails are placed into the sequence matters for this way of counting. However, for our actual problem, the order does not matter. Both of these sequences are the same sequence. Noting that each sequence is counted twice, we can see that the number of sequences with 2 tails is actually 45, exactly half of 90.

A similar argument can be used to compute the number of sequences with 3 tails, 4 tails and so on. The details get a little bit more complex, and we omit them from this introduction. Instead, we will use the EXCEL function COMBIN(N,K) – this counts the number of sequences

with exactly K heads amongst all sequences of length N. This function calculates the formula for “N choose K”, the number of different ways we can choose K positions from a total of N:

Calculations using this formula give the following results.

| #H | #SEQ | #H | #SEQ |
|----|------|----|------|
| 0 | 1 | 10 | 1 |
| 1 | 10 | 9 | 10 |
| 2 | 45 | 8 | 45 |
| 3 | 120 | 7 | 120 |
| 4 | 210 | 6 | 210 |
| 5 | 252 | 5 | 252 |

The Range of Normal Variation.

From these probabilities, we learn the following important lessons:

The probability of 5 Heads is 252/1024 or approximately 25%. The probability of 4 heads is 21% and 6 heads is also 21%

Before proceeding, note that there is a common misconception that if the probability of Heads is 50%, then 5 out of 10 flips will come out to be heads. This is not true for any individual trial of ten flips. It will hold on the average for a large number of trials. On any particular individual trial, the probability of seeing exactly 5 heads is 25% or 1 in 4. The probability of {4,6} combined is 42% so the chance of seeing one of these two numbers is greater than the chance of seeing 5 heads.

It is convenient to define a “range of normal variation” as being the set of values which have about 50% probability of being observed. With this definition, the chances of being within this set are about the same as the chances of being outside this set. If the number of heads is within the range of normal variation, then there is no reason to believe that the coin is not fair. Once we move outside the range of normal variation, then there is some evidence against the fairness of the coin.

With only 10 flips {4,5,6} is the range of normal variation of the flips. Outside this interval, we start getting evidence that the coin may not be fair. 3,7 provide some small evidence of bias. {2,8} provide more evidence. {1,9} provide even stronger evidence, and {0,10} provide the strongest possible evidence of bias in the coin. Nonetheless, it is important to note that no outcome provides conclusive evidence. Even if all 10 outcomes are heads, there is some chance that the coin may be fair, but there was a unusual streak of 10 heads appearing in a row. This event has probability 1 in 1000, but it is not impossible.

Central Events:

We will provide some technical terms which are used in connection with the problem under discussion. The NULL HYPOTHESIS is the assumption that the coin is fair. The median value of the number of heads is 5. This is a central value of the distribution, and always included within the range of normal variation. Any interval of values around the central value is called a central event. For example, {4,5,6} and {3,4,5,6,7,8} are both central events. Central events can also be asymmetric. For example, {4.5,6,7} is also a central event. For Binomials, probabilities are highest near the center and decline as we move away from the center. In this case, the following definition of the range of normal variation makes sense:

The Range of Normal Variation is the smallest Central Event which has probability greater than or equal to 50%.

For ten flips of a coin, {5} has probability 25% and does not qualify. {4,5,6} is the smallest symmetric central event which achieves probability 50% – in fact it has probability 67% which is quite a bit higher than 50%. This is because the small number flips lead to big jumps in probabilities as we move away from the center. When we look at larger number of flips this problem will disappear, and the range of normal variation will tend to have probability very close to 50%.

WHY do we name this central event the range of normal variation? If the random variable is just as likely to be within this range as it is to be outside it, then this is a very normal and usual value for the random variable. Such values provide no evidence against the null hypothesis. Once we go outside this range, then chances of being closer to the center are higher than 50% and chances of being further from the center are lower than 50%. This makes such events less likely under the null hypothesis. So such events provide some evidence against the null hypothesis.

Tail Events:

As we have discussed, unusual events provide evidence against the null hypothesis. The more unusual the event, the more the evidence against the null. Outcomes of 7,8,9, or 10 heads provide increasingly strong evidence against the null hypothesis that the coin is fair. {4,5,6} are within the range of normal variation, and hence provide no evidence against the null. The strength of the evidence against the null is often quantified as a “p-value”. We will provide a definition here, and note that the interpretation of p-values is quite complex.

Given any outcome of a sequence of flips of a fair coin, the (low) probability of the outcome measures how unusual the event is. The lower the probability, the more unusual the event. For any outcome K, the TAIL EVENT associated with K is the set of all outcomes which have probability less than or equal to K. This is the set of events which are even more unusual than K. Instead of measuring probabilities of individual events, it turns out to be useful to look at ALL events of probability less than or equal to the given event. The probability of the tail event associated with K is called the p-value of K. We now illustrate this p-value by showing how it is calculated.

Consider the outcome K=8 heads. From the table, the probability of this outcome is 45/1024. To create the tail event, we must consider all events with less than or equal probability. From the table, it is clear that the tail event for K=8 is {0,1,2,8,9,10}. These are all the events with probability less than or equal to 45/1024. Adding up the probabilities of all of these events gives us the p-value for K=8. The p-values for 8,9,10 are calculated below:

| Event | Tail Event | P-Value |
|-------|--------------|---------|
| 8 | 0,1,2,8,9,10 | 10.9% |
| 9 | 0,1,9,10 | 2.1% |
| 10 | 0,10 | 0.2% |

The smaller the p-value, the stronger the evidence against the null hypothesis. Sir Ronald Fisher popularized the rule that p-values of 5% or less were “significant” – that is, they provided significant evidence against the null. This is just a rule-of-thumb, and in any particular practical situation, many important considerations in addition to just the probabilities must be taken into account. The harm created by taking a rule of thumb, and converting it into a rigid scientific practice has been documented at book length by McCloskey et. al. in “[The Cult of Statistical Significance](#)“

A Second Example:

To get more practice with the concepts introduced in this lecture, we consider testing a coin for fairness by using 50 flips of the coin. We will just review the ideas introduced briefly. The first step is to calculate the probabilities of the outcomes. The middle value is 25, and the probabilities of outcomes very far from this are so close to zero that we do not need to tabulate them:

| #H | P(#H) | #H |
|----|-------|----|
| 15 | 0.20% | 35 |
| 16 | 0.44% | 34 |
| 17 | 0.87% | 33 |
| 18 | 1.60% | 32 |
| 19 | 2.70% | 31 |
| 20 | 4.19% | 30 |
| 21 | 5.98% | 29 |
| 22 | 7.88% | 28 |

| | | |
|----|--------|----|
| 23 | 9.60% | 27 |
| 24 | 10.80% | 26 |
| 25 | 11.23% | 25 |

Here we have used symmetry to compute probabilities of number of heads from 15 to 35. The probability of #H is equal to the probability of 50-#H, because $P(\#H = K) = P(\#T = N-K)$ these two are the same events. When #H=K then the number of tails must be N-K. Next due to symmetry between heads and tails, $P(\#H=S)=P(\#T=S)$. The same probabilities hold for Heads and Tails because these two are both 50% probability events. To find the range of normal variation, we compute the probabilities for the central events, going out from the center. Note that 25 by itself has only an 11.2% probability, much less than the 25% probability of 5 Heads in 10 trials. In general as the number of trials increases, the probability of the exact center decreases to 0. We consider an increasing band of central values around 25 and compute their probabilities by adding up the probabilities in the table above. This gives:

| Central Event | Probability |
|------------------|-------------|
| {25} | 11.23% |
| {24,25,26} | 32.82% |
| {23,24,25,26,27} | 52.01% |

According to our definition, the range of normal variation is from 23 to 27. These five outcomes have combined probability of 52% which is slightly higher than 50%. This is the first Central Event for which this holds. Values bigger than 27 or less than 23 provide SOME evidence against the null hypothesis that the coin is fair. How much evidence is measured by the p-value, which is computed in the table below:

| Event | Cum Prob | p-value |
|-------|----------|---------|
| 15 | 0.33% | 0.66% |
| 16 | 0.77% | 1.53% |
| 17 | 1.64% | 3.28% |
| 18 | 3.25% | 6.49% |
| 19 | 5.95% | 11.89% |
| 20 | 10.13% | 20.26% |

For each of the outcomes, the tail probability comes by adding up all even less likely outcomes. The Cumulative probability is $P(X \leq 15)$ which measures to total probability of all outcomes 0,1,2,...,15. In addition to these, we also have to take $50-15=35$ and all higher values,

since these have exactly equal probabilities. That is, 0=50, 1=49, and so on. The probabilities are symmetric from both ends of the table. Because of this symmetry it is sufficient to DOUBLE the cumulative probability, to capture the probabilities from the other side of the table. Thus gives us the table above. Using Fisher's rule of thumb, we see that 17 Heads or less, or 33 Heads or more have a p-value of 3.3% which is less than 5%. Thus, if we flip a fair coin 50 times and see less than 17 Heads, or more than 33 Heads, we should be suspicious about the null hypothesis that the coin is fair.

Concluding Remarks

On the surface, it does not appear like testing a coin for fairness has much to do with learning about association between two variables. But in fact, the topics are closely related, as we will see in the next part of this lecture. When we observe an association between two variables, one of the questions which must be resolved is: is this apparent association due purely to chance, or is it due to a genuine underlying real-world effect.

9E: Comparing Growth Rates

We will now turn to a detailed examination of an issue considered briefly before. Suppose we have a series of growth rates for two countries, such as Turkey and Thailand:

| | 1 993 | 1 994 | 1 995 | 1 996 | 1 997 | 1 998 | 1 999 | 1 000 | 2 001 | 2 002 |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| Turkey | 7 .8 | - 5.5 | 8 .5 | 7 .6 | 7 .6 | 2 .8 | - 3.7 | - .6 | 6 7.0 | - 7.0 |
| Thailand | 8 .9 | 7 .9 | 7 .9 | 5 .0 | - 3.2 | - 8.5 | - .6 | 5 .7 | 5 .7 | 2 .6 |

How do we decide which one has better performance, in terms of growth rates?

Previously, for a rough classification, we decided to use the MEDIAN growth rate of the two countries as a single indicator of growth performance. In this lecture, we will look at this problem in greater depth.

The question we want to think about is the following. Suppose that growth rates are determined by many random factors, and fluctuate, being sometimes high and sometimes low. Is it possible that both countries have growth rates determined by the SAME factors? That is, the differences that we see in the growth rates are just due to random variation, and not due to the superior performance of either country? The meaning of this question is easier to understand if we put it in the context of comparing two players scores in cricket.

Skill Versus Luck:

Suppose that in a given match, Javed scores 75 runs while Younis scores only 25. Can we conclude that Javed is the better batter? With only one match score, it does not seem like a good idea. A score depends on two factors: luck and skill. Skillful players will consistently score higher – on the average – than those with less skill. However, in any particular game, luck might

cause an upset, and the less skillful player might end up with a higher score. How can we distinguish between those two factors, and how can we come to a judgement that player X is more skillful than player Y? We now address this question. First consider the following ten scores for both players. Initially, we assume that both players played in the same matches.

| | | | | | | | | | | | | | |
|--------|---|---|---|----|----|---|---|---|---|---|---|---|---|
| Javed | 7 | 6 | 5 | 3 | 1 | 8 | 1 | 5 | 2 | 7 | 4 | 9 | 1 |
| d | 5 | 6 | 5 | 05 | 18 | 1 | 2 | 7 | 3 | 8 | | | |
| Younis | 2 | 5 | 3 | 7 | 7 | 4 | 1 | 4 | 7 | 6 | 6 | 6 | 1 |
| nis | 5 | 7 | 4 | 2 | 9 | 2 | 9 | 4 | 7 | | | | |

In each of the ten matches, Javed scored higher than Younis. Suppose the two players are equally capable. Then the chances of either one getting a higher score are 50-50 – either one is equally likely to get a higher score. In this case the chances of Javed scoring higher in each of 10 matches is like the chances of a fair coin coming up heads ten time in a row. This can be calculated to be 1 in 1024. This is a very low probability event. So it is hard to explain the observations on the basis that both players have equal skill. The observations are much more probable if Javed is the more skillful player. So the data leads to the conclusion – not definitive, but highly likely – that Javed is more skillful than Younis.

Some Philosophical Remarks: Conclusions we reach on basis of statistical analysis are almost NEVER certain. Numbers give us clues to reality; some clues are very strong while others are weak. But in all cases, final proof can only come from looking at the real world. The batting performance of Javed and Younis is only one – measurable and quantifiable – aspect of their skill at the game. An experienced observer who looks at the games played by both players can evaluate their skills in ways which cannot be captured by numbers and quantified. Indeed, all essential characteristics of human beings, like integrity, compassion, courage, intelligence, etc. are not measurable and not quantifiable. Unfortunately, in the West, the philosophy of positivism, and the worship of science, led to the opposite conviction: only observable, measurable and quantifiable aspects matter for science. This has led to serious distortions in the development fo statistics. It is not possible for us to go deeply into this topic here. We just make a few remarks to explain some key aspects of our Islamic approach to Real Statistics, which is very different from a conventional Fisherian approach.

Essential aspects of reality are always unobservable, and beyond the reach of numbers and measurements. The measurements that we can make are just rough approximations of a complex reality. If we reach some conclusion on the basis of these numbers, this is always a tentative conclusion, which may change if we learn some deeper aspects of reality, unknown to us at the time of analysis. Just to illustrate this, we provide some examples of how the same numbers which suggest that Javed is a superior player may lead to a different conclusion,. But there can be cases where an expert observer might say, by looking at the two players, that Younis is the superior player, but happened to have a streak of bad luck on these games. Or he might be able to offer some deeper explanation of the numbers on the basis of factors we have not considered.

Some Cases where the same numbers lead to different conclusions:

Next, consider the same ten numbers re-arranged as follows:

| | | | | | | | | | | | | |
|--|--------|-----|----|----|----|----|----|----|----|----|----|---|
| | Javed | 105 | 13 | 98 | 85 | 76 | 61 | 57 | 45 | 32 | 28 | 1 |
| | Younis | 6 | 9 | 5 | 4 | 2 | 4 | 7 | 7 | 2 | 9 | 7 |

The set of 10 scores for both players is exactly the same, but they have been arranged in decreasing order for Javed, and in increasing order for Younis. This leads to higher scores for Javed in the first 6 games, and higher scores for Younis in the last 4 games. Can we still conclude that Javed is the better batter? Probably not – there seems to be some other process going on through time that we do not know about. Perhaps Younis is improving while Javed is going down. If this is so, then Younis would be the better batter by the end of the season.

Next consider a situation where both players achieved the same 10 scores as listed above, but they played in different matches. Can we still conclude that Javed is the better player? What is the difference between the same match situation and the different match situation? In the same match, a lot of possible confounding factors are eliminated. The conditions under which both players played are “roughly” the same. Of course it is possible that one player always faced a tougher bowler (by chance), or that one was further down the list of batsmen and therefore did not get to bat for as long. So while there may be factors other than skill which caused the difference in performance, the possibility is minimized when scores are compared in the same matches. But if we compare scores when the two players played in different matches, then a large number of possibilities for confounding arise. As a simple example, suppose that Younis played these ten games in international competition, while Javed’s scores are in domestic competitions, which are a lot easier. That would obviously make the scores incomparable, and we could say nothing about the relative skill of the two players.

Back to Growth Rates:

Now we come back to the original problem under discussion. Given the record of two countries performance on growth rates over the past 60 years, we would like to decide if one of the two is SIGNIFICANTLY better than the other. Perhaps it is just chance fluctuations which lead to improved performance, or perhaps one of the countries has significantly superior macroeconomic performance due to real factors which we should investigate? We can consider the growth rates as the scores of the two countries on 60 matches, and ask if one of the two is significantly better. The NULL hypothesis is that there is no significant difference in performance of the two countries. If this null hypothesis is true, then either one of the two countries can have the higher growth rate with equal probability. In other words, the probability that country A has higher growth than country B is exactly 50%. So we can check this null

hypothesis by looking at which of the two countries has higher growth rate and treating this as a sequence of flips of a fair coin. If the number of times country A is higher is around 50%, then the null hypothesis is confirmed. If the set of wins for A is significantly more than 50% than A has better performance. We now do some data analysis based on this conceptual framework.

Only a few countries, about 18, have nearly complete data sets on growth rates, starting from 1960. We list these countries, by their WDI three letter codes, below. The full names are:

Argentina Madagascar Greece South Africa Australia Algeria Congo, Rep. Mexico Ecuador
 Guatemala Chile Peru Philippines Thailand India Malaysia Korea, Rep. Singapore, in the same sequence as listed in the table below. These countries have been ranked by median growth rates in increasing order. Countries lower down in the table have higher median growth rates. The question of interest to us is: Is the difference significant (that is, due to genuine real-world factors) or could it be due to pure chance? We have developed a methodology to answer this question, and we will apply it to get an initial and tentative answer to this question. A final answer can only be obtained after looking more deeply at the development and growth factors for the countries in question

| | C ode | 1 961 | 1 962 | 1 963 | 1 964 | 1 5:'15 | 6 016 | 2 017 | 2 018 | 2 019 | 2 020 | ME DIAN |
|----|------------|------------|-----------|-----------|----------|------------|-----------|-----------|-----------|----------|----------|------------|
| RG | A .65 | 5 1.01 | - 5.31 | - 0.31 | 1 4 | .. 2.25 | - .23 | 2 3.57 | - 2.35 | - 2.3 | - 3 | 2.2 |
| DG | M .12 | 2 .23 | 2 1.01 | - .01 | 4 .. | .. .60 | 3 .02 | 5 .57 | 4 .46 | 4 .42 | 4 2 | 2.3 |
| RC | C 3.32 | 1 .31 | 0 1.97 | 1 .54 | 9 .. | .. 0.32 | - .70 | 1 .20 | 1 .. | 1 3 | 1 3 | 2.5 |
| AF | Z .56 | 3 .96 | 6 .70 | 7 .80 | 7 .. | .. .09 | 0 .14 | 1 .62 | 0 .50 | 0 .7 | 0 7 | 2.9 |
| US | A .24 | 2 .64 | 1 .69 | 5 .23 | 7 .. | .. .49 | 2 .93 | 1 .50 | 2 .66 | 1 8 | 1 8 | 3.1 |
| ZA | L 13.54 | - 19.84 | - 4.14 | 3 .80 | 5 .. | .. .76 | 5 .46 | 6 .. | 0 0 | 0 3 | 0 3 | 3.7 |
| OG | C .96 | 7 .38 | 5 3.83 | - .82 | 3 .. | .. 9.07 | - 6.14 | - 2.72 | - 0.34 | - 2 | - 2 | 3.8 |
| EX | M .13 | 5 .83 | 4 .07 | 8 1.45 | 1 .. | .. .74 | 2 .22 | 2 .06 | 2 0.55 | 2 4 | 2 4 | 3.9 |
| CU | E .75 | 4 .95 | 4 .77 | 2 .57 | 7 .. | .. 1.30 | 1 .91 | 1 .85 | 0 0.22 | 0 7 | 0 7 | 4.0 |

| | C | 4 | 3 | 9 | 4 | .. | 2 | 3 | 3 | 4 | 4.0 |
|----|----------|-----------|-----------|-----------|---|-----------|----------|----------|----------|--------|-----|
| TM | .12 | .67 | .45 | .38 | | .91 | .09 | .25 | .27 | .7 | |
| HL | C .43 | 5 .97 | 3 .86 | 5 .47 | 2 | .. .57 | 1 .04 | 0 .80 | 3 .36 | 1 1 | 4.3 |
| ER | P .75 | 7 0.81 | 1 .43 | 4 .27 | 6 | .. .17 | 3 .95 | 1 .64 | 3 .44 | 2 2 | 4.3 |
| HL | P .40 | 6 .18 | 5 .05 | 7 .38 | 3 | .. .76 | 6 .82 | 6 .88 | 5 .29 | 5 3 | 5.1 |
| HA | T .38 | 5 .56 | 7 .02 | 8 .84 | 6 | .. .93 | 3 .25 | 4 .70 | 3 .66 | 3 3 | 5.6 |
| ND | I .61 | 3 .92 | 2 .04 | 6 .40 | 7 | .. .30 | 7 .08 | 8 .13 | 6 .02 | 9 9 | 5.7 |
| YS | M .19 | 9 .54 | 7 .53 | 7 .78 | 4 | .. .43 | 4 .66 | 5 .95 | 3 .06 | 5 9 | 6.8 |
| OR | K .35 | 7 .81 | 3 .67 | 8 .72 | 9 | .. .00 | 3 .20 | 3 .51 | 2 .68 | 2 7 | 7.6 |
| GP | S .14 | 8 .49 | 7 0.68 | 1 2.23 | - | .. .70 | 3 .74 | 2 .89 | 0 .22 | 1 5 | 7.8 |

In this list of 18 countries, we can make $(18 \times 17)/2 = 306/2 = 153$ pairwise comparisons of two countries. To illustrate the methodology, let us compare the first country Argentina with the 17 other countries. We will look at the process of comparison for Argentina (ARG) and Mexico (MEX) in detail, to explain how it is done. The first step is to compare the growth rates for the 59 years of data on growth rates for the two countries. In the EXCEL spreadsheet L7D SigDif GrRate (Significant Differences in Growth Rates), the sheet named 18 countries replicates the original WDI data set in Rows 3 to 20 for the 18 countries for which we have growth rates starting from 1960. The required calculations are done in rows 23 to 39 – these provide us with comparisons of ARG with all other countries. Let us look at ROW 29 which compares ARGentina to MEXico.

First Step: Column F has data on growth rates for 1960, with F3 being for ARG and F10 being for MEX. In cell F29, we enter “=IF(F3>F10,1,0)”. This EXCEL command compares F3 to F10 and enters 1 in cell F29 if F3 is greater than F10. Otherwise, if F3 \leq F10, then the value of 0 is entered in the cell F29. Now we replicate this entry in ROW 29 in all columns from F to BL – each column contains data for one year, going from 1960 to 2019. Thus, we get a sequence of 1’s and 0’s from F29 to BL29, where the 1’s capture the years for which ARG had higher growth rate than MEX. The 0’s are entered in the years where the reverse was true, and MEX had the higher growth rate.

Second Step: Next we COUNT the number of 1's in this sequence of 59 years of data. This can be done by putting the command “=SUM(F29:BL29)” in cell E29. This yields the value 24. In 24 years, ARG had the higher growth rate, while in the remaining 35 years, MEX had the higher growth rate.

Third Step: 24 is a little less than 50% of 59, or 29.5. Is the difference significant? To find out, let us compute the p-value. The tail event associated with 24 is all values from 0 to 24 and also all values from $59-24=35$ to 59. The EXCEL command =BINOMDIST(24,59,0.5,true) computes the cumulative probability of all values from 0 to 24 when a fair coin is flipped 59 times. This comes out to 9.63%. By symmetry, the probability of all values from 35 to 59 is exactly equal to this, and is also 9.63%. So, to get the p-value, we can just DOUBLE this value to get 19.25%. This is listed on the EXCEL spreadsheet. This p-value is too large. The interpretation is that a deviation of a fair coin from the central values of 29 and 30 by 5 units in either direction (positive or negative) will happen about 20% of the time. So this is not very unusual. This means that there is not much evidence in favor of the idea that MEX had significantly better performance than ARG. Economists of Argentina should not necessarily look to Mexico for guidance in terms of how to manage their macroeconomic policies.

We can replicate these calculations in order to compare growth performance of Argentina with all of the other 17 countries on which we have (nearly) complete data on growth rates. Wins counts the number of times ARG has higher growth rate, in the 59 years for which data is available.

| Country | p-val | Wins | Country | p-val | Wins |
|------------|---------|------|------------|--------|------|
| MDG | 100.00% | 29 | CHL | 6.74% | 22 |
| GRC | 79.48% | 28 | PER | 11.75% | 23 |
| ZAF | 60.29% | 27 | PHL | 3.63% | 21 |
| AUS | 60.29% | 27 | THA | 6.74% | 22 |
| DZA | 100.00% | 30 | IND | 3.63% | 21 |
| COG | 60.29% | 27 | MYS | 0.38% | 18 |
| MEX | 19.25% | 24 | KOR | 0.15% | 17 |
| ECU | 79.48% | 28 | SGP | 0.02% | 15 |
| GTM | 6.74% | 22 | | | |

For 9 countries (Madagascar Greece South Africa Australia Algeria Congo, Rep. Mexico Ecuador Peru the p-values are very high, and there is no evidence of superior growth performance at all. Among these 9, the lowest p-value is 11.75% for Peru. According to reasonable Fisherian convention, this is too high to signal a genuinely superior growth

performance. However, additional evidence of other kinds may over-ride this conclusion. For the remaining countries, Guatemala, Chile, and Thailand have p-values of 6.74, which is slightly above the Fisher cutoff of 5%. Since this rule should not be applied mechanically, it may be worth considering these three countries as possibly being superior performers in terms of growth. The p-values of 3.6% for Philippines and India are significant, and suggest that these two countries had genuinely superior growth performance. Malaysia, Korea, and Singapore have p-value of less than 1%, which is highly significant. The data provides strong evidence that the superior growth rates of these countries are not purely due to chance. Rather, some genuine real-world factors – better macroeconomic policies or more favorable environment – are responsible for this superior economic performance.

Concluding Remarks:

What do we learn from this analysis? Some of the key lessons, which differ from the conventional (Fisherian and positivist) approach, can be summarized as follows:

1. Our main goal is to learn about the real world. Numbers are only indicators of real world events, and almost never the goal of our analysis. What we want to learn about the world cannot usually be captured by the numbers.
2. Numbers can provide CLUES to real world processes of importance. Learning occurs by FOLLOWING these clues, in order to study the real world in greater depth. This will usually involve expanding our study to qualitative, unmeasurable, and even unobservable phenomena.
3. For example, the numbers might show that player X is more skillful than player Y. But the analysis of the nature of this skill will involve deeper study of the game and the player.
4. Our study of growth rates shows that apparent differences between Argentina and Mexico are likely to be due purely to chance. This means that the question of “Why is the median growth rate of Mexico (3.94%) higher than the Argentina’s 2.23%” is NOT a good question to ask – the answer may be that it is purely due to chance. We will discover nothing if we study factors responsible for the higher median growth rate of Mexico. On the other hand, it seems clear that Malaysia did have higher growth rates than Argentina. If we study factors which caused Malaysia to have higher growth rates, we expect to find something. The numbers point to a phenomena which requires explanation, since it is not due purely to chance. However, the numbers themselves contain NO INFORMATION about what that phenomenon might be.

10: Some Applications

Chapter 10 blurb should described the 5 or 6 applications we make in this chapter.

10A Assessing the Salk Vaccine for Polio

Although concepts discussed in this lecture are elementary, they are unfamiliar and require some work to digest. It is recommended to do the reading (PDF attached) first, then watch the video, and THEN read this textual material for complete understanding.

At date of this writing 15 Nov 2020, COVID cases are on the rise. While people anxiously await development of a vaccine for COVID-19, few are aware of the statistical aspects of tests required to assess a vaccine for safety and effectiveness. This lecture is about the tests which were carried out for the Salk Vaccine against Polio in the 1960's, to explain some basic issues related to this process. There are more complex issues involved in assessing whether or not the vaccine has any undesirable side effects. These are not dealt with in the present lecture.

How do we know if a vaccine against Polio (or COVID) is effective? A little thinking shows that this may not be very easy. We give the vaccine to someone, and observe that he/she does not get polio – does this mean that the vaccine is effective? Of course not – he/she may not have gotten polio anyway, without the vaccine. Unlike COVID, which is highly contagious, Polio is contagious but on a smaller scale. In general, without vaccine, the rate at which children contract polio is less than 1 in 1000. This means that if we give the vaccine to a 1000 people, and none of them gets polio, we cannot conclude that the vaccine was effective. Only 1 person would have gotten polio anyhow, and the difference between 1 per 1000 and 0 per 1000 is so small that it could be due purely to chance. So, to begin with, we need a large sample size to test the vaccine for effectiveness.

The lecture studies the setup and outcomes of two studies: NFIP and RCDB. Both have huge sample sizes involving around 200,000 children who were given the vaccine. A key lesson in testing vaccine is the necessity of having a comparable control group. There must be two groups – treatment and control – which are well matched on all factors relevant to the disease. The NFIP study took Grade 2 as the treatment group and Grades 1 & 3 as the control group. Roughly speaking, this seems like it should produce a good match between treatment and control, but there are two problems with this design. One is that Polio is a contagious disease. This means that if one student in class catches, many others in the class are also likely to catch it. In the NFIP experiment, the vaccine was given to 225,000 children in the 2nd grade, and 56 of them developed polio. As we will see, the Salk Vaccine *is effective*. So we must abandon the common misconception that if the vaccine is effective than NO ONE would have developed polio. Even an effective vaccine can fail, for many different reasons. To know whether or not the vaccine worked, we must compare with an *equivalent* control group, which did not take the vaccine. A huge number of problems arise in ensuring the EQUIVALENCE of the control and treatment groups. SOME of these problems are discussed in this lecture.

The only way to know if the Salk Vaccine was effective was to COMPARE the rates of Polio in the Treatment Group (Grade 2) and the Control Group (Grades 1 & 3). If the polio rates in the control group were SIGNIFICANTLY higher, then we should learn that the difference is not due to chance alone, it must be due to SOME OTHER RELEVANT FACTOR which is different in the treatment and control group. One OBVIOUS relevant factor is the Treatment

itself – everyone in the treatment group took the Salk Vaccine drops, while everyone in the control group did not. But there may be OTHER relevant factors which caused the difference. Such factors are known as CONFOUNDING factors.

Here is a simple illustrative example of a confounding factor which could cause problems for the NFIP experiment. Suppose that the risk of declines rapidly with age. It is 100 per 100,000 at age 5, 60 per 100,000 at age 6, and 50 per 100,000 at age 7. Also suppose the 5,6, and 7 are the ages of children in grades 1, 2, and 3. Then age becomes a confounding factor in the NFIP study. If Salk Vaccine is completely ineffective, Grade 2 rates will be 60 per 100,000 which is smaller than 75 per 100,000 the average rate of grades 1 & 3. Thus it will seem as if the Vaccine prevents about 15 cases per 100,000 of polio when in fact this effect is purely due to the AGE factor. This is called confounding. There could be other UNKNOWN factors which are different between Grades 1 and 3 and the Grades 2 chosen for the NFIP study. It is important to note that differences which do not affect outcomes are not relevant here. For example, suppose that polio is gender-unbiased – it affects males and females equally. Also, the effectiveness of the vaccine is not affected by gender. Then it does not matter if the treatment and control group are not matched on gender. We will get correct results regarding the effectiveness of the vaccine even if the treatment group is all females and the control group is all males.

Empirical evidence shows that AGE was not a confounding factor in the NFIP design, even though it could have been. The differences in polio rates among the ages 5 to 8 for grades 1,2, and 3 were not large enough to matter. But there was a subtle and unexpected confounding factor. That was the problem of CONSENT. Since this was an experiment, and there was a potential that the Salk Vaccine might have harmful side-effects, it was necessary to get consent of the parents before giving the vaccine to the children in grade 2. Some of the parents allowed their children to take the experimental vaccine against polio. Other parents refused. So the Treatment Group in the 2nd Grade consisted not of ALL 2nd grade children but only of those children whose parents consented. The question arises: does this MATTER? Initially, it does not seem that it should make any difference. Whether or not parents consent to the vaccine should not matter for the effectiveness of the vaccine. However, some experimental data showed that this was not the case. There were systematic differences in the consenters and those who did not consent. Generally speaking, higher income families were less suspicious of medical interventions and readily gave consent. Low income families tended more often to not consent. This fact was noticed when demographic characteristics of consenters and non-consenters were studied. Again, this should not matter – income should not play a role in the effectiveness of the vaccine. However, it turns out the income does matter for chances of getting polio. This is because hygiene is relevant for polio. In poor families with poor hygiene, there were more chances of low-level exposure to polio which would create some immunity against polio. On the other hand, in high income families, children would be at higher risk for polio because good hygiene would prevent previous exposure and create lack of immunity.

This was one of the defects of the NFIP design – the control group and the treatment group were not matched with respect to income. The treatment group consisted of children whose parents consented to the vaccine and this group had higher average income. The control

group of Grades 1 and 3 were not asked for consent, and so the children of parents who would not have consented were also included in the control group. The higher income of the treatment group led to greater vulnerability to polio in the treatment group. The lower average income in the control group – which was a mixture of consenters and non-consenters – led to less vulnerability to polio. The data shows this difference. The control group of consenters in the RCDB trials received no vaccine and had polio rates of 71 per 100,000. The group of non-consenters in RCDB study had lower polio rates of 45 per 100,000. This is a significant difference. Both the control group of consenters, and the no-consent group did not receive the vaccine but the rate of polio was significantly higher in the consenters who did not receive vaccine.

There is one more problem with the NFIP design mentioned by [Freedman \(2007\)](#). This comes from the fact that polio is a contagious disease. To take an extreme example, suppose that in one of the classrooms selected for the study, one student manages to infect the whole class of 30 students. Given that total polio cases are fairly small in number, just one case like this would make a huge difference to the outcome. Depending on whether the contagion occurred in the control group or the treatment group, it could bias the study for or against the vaccine.

So how can we solve these problems, and create a design which will give us better results in terms of evaluating the vaccine?

Randomization: A very important and ingenious technique developed by Fisher involves assigning subjects to treatment and control group at random. This ensures a rough match between the treatment and control group on confounding factors. To be more explicit, suppose that there is a confounding factor F which we do not know about and which cannot be observed (like some gene which affects polio outcomes). Suppose that 87% of the population carries this gene F which provides protection against polio, and 13% does not. If we assign subjects to Treatment and Control group at random from this population, and if the sample size is sufficiently large, then both Treatment and Control group will also have approximately 87% subjects with gene F. Thus the two groups will be matched, and any difference in outcomes for the two groups will be due to the vaccine and not due to the confounding factor F.

Consent: As discussed earlier, the consent and no-consent groups are different with respect to income, and this affects their responses to polio. Randomization cannot help us here because if we choose equal proportions of no-consenters for the treatment group, we still cannot forcibly give them the treatment. The only solution is to leave the no-consent group out of the experiment. This means that we first get consent from parents to participate in experiments on the vaccine. Then we choose the treatment and control group at random from the consenting group. In this way, the two groups will be matched in terms of consent – both groups will be consenters.

Contagion: With this design, some of the consenting students from the same classes will be assigned to treatment and some to control at random. Thus both groups will share the same environment and will be subject to the same contagion risks. Any difference in polio outcomes

would therefore be due to the vaccine, and not due to one group being exposed to contagious environment while the other is not.

POST RANDOMIZATION PROBLEMS: Even though randomization is known as the gold standard, the best way to design an experiment, there are problems which cannot be resolved by randomization. One of these is consent, as already discussed above. Suppose we wish to study effects of smoking on cancer. We cannot randomly assign subjects to Treatment and Control groups and ask one group to smoke for ten years, while asking the other group to quit smoking, for the sake of our experiment. We can only observe what happens to smokers and non-smokers and attempt to match the treatment and control groups by some OTHER method, because randomization cannot be used. This is called an OBSERVATIONAL study – we will discuss this type of study later, and see how this makes it difficult to avoid confounding. Secondly, there are differences created by the treatment itself. There is a psychological phenomenon known as the “*placebo*” effect. If we tell the patient that we are giving him a drug to treat him, and give him a blank pill which contains no medicine, the patients often get better. The knowledge that a treatment is being given creates a confidence in the patient which leads to healing. Experiments are carried out “BLIND” to resolve this problem.

Blinding the Subjects: We can resolve problems created by the Placebo effect by not letting the subjects know whether they are in the treatment group or the control group; going even further, we may not let them know that there are two groups. To do this, everyone is given medication, but the medication is just water or a dummy pill for the control group, while it contains medicine for the treatment group. This ensures that all subjects are in the same boat with respect to their knowledge about receiving treatment. This should eliminate the placebo effect, although there are cases where this does not work. The patients learn, against the will of the experimenters, about which group they belong to, and whether or not they are receiving treatment. This can cause confounding problems.

Blinding the Experimenters: In a double blind study, the doctors who give the treatment to the patients do not know which patients are receiving the vaccine and which are receiving dummy tablets without any medication. Why do this? How can it matter whether or not the doctors know who is being treated? In the case of polio it can matter because the doctors have to decide which children have contracted polio, and which children did not. It turns out the polio has many different forms and appearances, and is not easy to diagnose. In borderline cases, the doctor could be influenced by his knowledge of whether or not the child has had the vaccine. If a child is in the treatment group, the doctor might be inclined to judge him polio-free when the symptoms are ambiguous and not clear-cut. Conversely, if the child is in the control group, the doctor might think he has polio with the same set of ambiguous symptoms.

10B Discovering Causes & Differentiating them from Correlates

In this lecture 8B, we study some real world cases which show how causes were discovered. One of the key ideas is that Correlation is NOT Causation. That is, we cannot learn about causes just by looking at the patterns of association within the observed data. Discovering

causes requires going beyond the observations to the underlying real world structures and mechanisms which generate these observations.

In fact, one of the Main Problems we face in statistics and econometrics is “How to Discover Causes when all we observe is Correlations?” It will surprise students to know that conventional statistics has NO ANSWER to this problem. The standard techniques, currently in use in statistics and econometrics use the following Naïve Approach. We start with a GUESS that X causes Y. We calculate the correlation between the two, or else run a regression of Y on X and other suitable variables. If the correlation between Y and X comes out significant, we conclude the X is a cause of Y. This methodology is seriously flawed. Some of the problems are the following:

1. The correlation may be due to Reverse Causality: Y is a cause of X and not the other way around. Correlations do not allow us to distinguish the two cases.
2. Common Cause: If there is a variable Z such that Z causes Y and also causes X, this will lead to association between X and Y. However, in this case there is no direct causal relationship between X and Y.

There are many other causal structures which could lead to an appearance of correlation between X and Y without any causal connections between the two. Since discovery of causation is of central importance for policy, it is a puzzle why statistics and econometrics has such seriously deficient methodologies for the discovery of causes.

The reasons why Statistics is Blind to Causality emerge from a Long and Complex History. Chapter 2 of “The Book of Why” describes the bizarre story of how statistics inflicted causal blindness upon itself. This has to do with early mistakes and confusions by Galton, Pearson, Fisher, the founding fathers of modern statistics. In fact, these mistakes were due to the effects of the philosophy of Logical Positivism. According to this (mistaken) philosophy, scientific theories should not go beyond the observables. Since causes are never observable, prohibitions on discussing unobservables made it impossible to talk about causality. Loss of language and terminology for causation led to hopelessly bad methodologies, which are continue to be in use in mainstream textbooks, even though substantial progress has been made in the past few decades.

In this lecture, we will look at three different real-world case studies relating to the discovery of causes. The first of these is the LEAPS Survey, the second is about the discovery of the causes of Puerperal Fever, while the third shows the differences between randomized controlled trial and observational studies.

Case 1: Why are enrolments higher for male children compared to females, in rural Punjab?

World Bank study entitled: Learning and Educational Achievements in Punjab Schools (LEAPS): Insights to inform the education policy debate, published February 20, 2007, authored by Tahir Andrabi, Jishnu Das, Asim Ijaz Khwaja, Tara Vishwanath, Tristan Zajonc

We will focus on just ONE insight into causes which was discovered by the LEAPS Survey. The Survey confirmed the commonplace observation that enrollments of female children in schools are LOWER than those of males. The question is: What is the CAUSE of this differential? Given the culture and traditions of Pakistan, there is an obvious Hypothesis: "**H1: Social Norms favor educating boys, and are against educating girls.**" If this is true, then campaigns to change mindsets are needed. It is very important to note that enrollment data can tell us nothing about this. Traditional methodology in both econometrics and statistics would proceed by putting forth the hypothesis H1, noting the lower enrolment of girls, and taking this as a PROOF of the null hypothesis. As the survey shows, this is the wrong conclusion.

One of the goals of the survey was to discover if Parents invest equally in education of all children, or whether they favor the brighter children. Survey questions lead to the result that parents invest the most money in education of brightest child – regardless of gender! Thus the initial Hypothesis H1 is false. The question remains: How to explain greater enrolment of male children? We can think up a New Hypothesis: "**H2: Parents THINK male children are brighter than females**". If parents are biased in this way, then male enrolments would be higher, even though parents think themselves to be impartial between genders. However, the Survey Data Rejects this hypothesis H2 as well. The survey gathered three different sources of evaluation for children – test scores at schools, teachers evaluations, and parents evaluations of the children. All three were in good correspondence with each other. This was a bit SURPRISING, since often parents are IGNORANT about the subjects being taught. NONETHELESS, they can accurately evaluate schools and teachers who do well at educating their children. They are also able to judge the relative capabilities of different children accurately.

To make progress on discovering the causes, we need to distinguish between intentions and actions. Note that Positivists REJECT the distinction. The hidden intention is expressed by the action. However, conversations between the survey team and parents led to the realization that Parents often say we would like to educate our girl, who is brightest, but there are no schools nearby. This type of informal and qualitative evidence led to the realization that distance to the nearest school is a key variable. The study shows that parents are willing to send male children further away to schools, but are much less willing to send daughters to distant schools. This insight emerges from study of real world, and is clearly not part of original data.

Note how important finding the cause is for appropriate Policy. Without any deeper study, many have actually concluded that wrong social norms are the cause of the problem, and are accordingly working on campaigns to educate parents about the value of educating girls. According to the survey, parents already know this, and hence this is wasted effort. If hypothesis H2 is correct, then we need to inform parents about educational abilities of their children. However, this also is not need. The third hypothesis informs us that to increase female enrollment, we must lower distance to schools. This may involve building a lot more schools for females, and may not be feasible financially. Alternatively, we need to invest in providing secure transport to school for females, and work on alleviating anxieties of parents in this regard.

Case 2: Ignaz Semmelweiss discovers cause of Puerperal Fever

When IS arrived at the Vienna General Hospital in 1846, he learned that there were two essentially identical clinics A & B located next to each other, which had dramatically different rates of mortality. Both were maternity clinics meant for delivery of children. Women arriving were admitted to A or B on alternate days, so the two sets of patients should have been the same. The staff, methods of treatment, doctors, were the same at both clinics. Then why was Mortality of Women almost double in Clinic A? It is important to note that in mid 19th Century, there was no concept of germs. Underlying theories of disease had to do with Miasma: Bad Air, combined with weakness in patients, leads to disease. But, IS wondered, “How can there be a DIFFERENTIAL in Clinics A & B, with alternating admissions?” Both clinics shared similar environment, and so Miasma+Patient Weakness should lead to same outcomes in both clinics. Many other hypotheses about the possible cause of the difference were studied and rejected: Sunshine, Diet, Use of Hospital Linen, and others. None of these could explain the dramatic differential in mortality due to Puerperal Fever in females after delivery.

An accidental event led IS to formulate a new hypothesis about the cause of this differential. A colleague Kolletschka was accidentally cut with a knife used in autopsy. He then developed symptoms like puerperal fever and died. This led IS to the conclusion that “cadaveric materials” – that is, some parts of the dead body (cadaver) – were transferred into the bloodstream of Kolletschka. Guessing that this is the cause led IS to examine how/why women in clinic A were exposed to possibility of infection from dead bodies, while those in clinic B were not. Once the hypothesis about the cause was formulated, it was easy to find a lot of confirming evidence for it. Clinic A was dedicated to training male medical students, while clinic B trained females to be midwives. Medical Students in clinic A routinely do autopsies on cadavers, then go on to examine pregnant females. Without germ theory, there was no emphasis on hygiene, and students would often examine females immediately after coming from an autopsy, without washing hands. In contrast, students training to be midwives had no contacts with cadavers, and did not do autopsies.

What seems obvious to us today, was revolutionary in 1850. The ideas that transfer of infectious materials from cadavers was contrary to accepted medical theories. Nonetheless, IS was able to find many Confirming Observations for his cadaveric theory of puerperal fever. He learnt that prior to 1823, the hospital director had been against the use of cadavers for medical training, and hence not autopsies were performing. After change of director in 1823, cadavers came into use for autopsies and medical training for students. At the same time, mortality rose from 125 to 500 per 100,000 in both clinics. The two clinics were separated into clinic A for male students and clinic B for midwives in 1840. It was only after 1840 that the differential appeared, where mortality rates in clinic A were twice as high as clinic B. Another evidence confirming the cadaveric hypothesis was that Puerperal fever occurred in rows in male Clinic; this was because male students entered and examine female patients in sequence, infecting them in rows. In contrast, in clinic B, cases occurred at random, here and there, without apparent sequencing.

Based on this diagnosis, Semmelweis came up with a simple solution: students should wash hands with Chloride of Lime after autopsy. This simple remedy dramatically reduced cases

of puerperal fever in the clinic A, apparently confirming the validity of IS hypothesis about the cause. However, later developments showed that the cadaveric hypothesis was not quite right. In one case, even though all students washed their hands, all women in a ward with one female with cancer got infected with puerperal fever. The female was the first in line to be examined. This, and many other cases, showed IS that origins of disease were not caused only by cadavers. Accordingly, he came up with a Modified Hypothesis: "Decaying organic matter causes infections". This is essentially correct and as close as you could get to the modern germ theory in the 19th Century. If the cadaveric hypothesis is true, then it would be enough for students to wash hands after autopsies or contact with cadavers. However, if the modified hypothesis is true, then students should wash hand prior to every examination. The second is the correct policy, in light of our modern knowledge of germ theory. Note how correct policy changes drastically as we learn more about the correct causes. Note also that even incorrect causal theories can improve policies. Note also the dramatic difference it makes to learn the true causes of disease in terms of saving lives.

Case 3: Difference between Randomized & Observational Studies

In the 1960's, heart attacks emerged as a major cause of mortality in the USA. These appeared to be linked to cholesterol levels in the blood. Many experimental drugs to lower cholesterol and reduce heart attacks were developed. One of these drugs was CLOFIBRATE. The Coronary Drug Project was a randomized, controlled double-blind experiment, whose objective was to evaluate five drugs for the prevention of heart attacks. The subjects were middle-aged men with heart trouble. Of the 8,341 subjects, 5,552 were assigned at random to the drug groups and 2,789 to the control group. The drugs and the placebo (lactose) were administered in identical capsules. The patients were followed for 5 years. The results of this trial are summarized below:

Table 1. The clofibrate trial. Numbers of subjects, and percentages who died during 5 years of followup. Adherers take 80% or more of prescription.

| | <i>Clofibrate</i> | <i>Placebo</i> | | |
|--------------|-------------------|----------------|---------------|---------------|
| | <i>Number</i> | <i>Deaths</i> | <i>Number</i> | <i>Deaths</i> |
| Adherers | 708 | 15% | 1,813 | 15% |
| Non-adherers | 357 | 25% | 882 | 28% |
| Total group | 1,103 | 20% | 2,789 | 21% |

Note: Data on adherence missing for 38 subjects in the clofibrate group and 94 in the placebo group. Deaths from all causes.

Source: The Coronary Drug Project Research Group, "Influence of adherence to treatment and response of cholesterol on mortality in the Coronary Drug Project," *New England Journal of Medicine* vol. 303 (1980) pp. 1038-41.

Does Clofibrate Save Lives? The table say NO. There are 21% deaths in control group and 20% deaths in treatment group. Currently there are 220 deaths in the treatment group and only 11 more would bring it up to 21%. Such a small difference could easily occur by chance. However, the doctors running the trial were puzzled by this result. This is because Clofibrate did lower cholesterol – why did it not lower the mortality from heart attacks? The experimenters thought that perhaps the mortality rates were high because some of the patients were not taking the drugs. They were able to divide the patients into two groups. Group A – the Adherers – took more than 80% of the drugs. Group NA – the Non-Adherers – took the drug less than 80% of the time. The table above shows that the adherers have only a 15% mortality rate, while the non-adherers have a 25% mortality. This is a significant difference. Can we conclude that clofibrate DOES make a difference, provided that you take the drug regularly? To answer this question, we must look at the treatment and control group in this NEW comparison.

Here Group A is the treatment group, the patients who took the Clofibrate regularly, more than 80% of the time. Group NA is the control group, which took the drugs irregularly, less than 80% of the time. These groups have NOT been randomly chosen, and so there is no guarantee that the two groups are matched. One obvious difference between groups A and NA is that regular taking of drugs indicates health-consciousness. Perhaps all members of group A are careful about their diet, exercise, as well as health. Since group NA did not take the drug regularly, perhaps they are not so health-conscious. Perhaps they are lax about exercise, diet and health in general. If this is so, Health-Consciousness is a Confounding Variable. The difference in the mortality rates between Adherers and Non-Adherers could be due to Health-Consciousness rather than the Clofibrate. How can we tell whether or not this is the case? One way is to look at the control group. The control group did not get the Clofibrate. Instead they got dummy pills which looked like Clofibrate, but they contained lactose, which has no effect on heart attacks. In the control

group also there were adherers and non-adherers. If reduced mortality for Adherers is due to Clofibrate, then the Adherers who were taking dummy pills should NOT have lower mortality – they did not get the clofibrate. However, the table above shows that Groups A and NA have a huge difference in mortality rates of 15% to 28%. This difference cannot be due to clofibrate because no one in the control group got the drug. Therefore the difference between groups A and NA is caused by the health-consciousness, and NOT by the clofibrate.

The point of this discussion is that whereas we can trust that randomized experiments match the control and treatment groups, this cannot be assumed when the two groups are not randomly chosen. In the second case, which is called an observational study, there may be other differences, called confounding factors, which affect the outcomes. In the clofibrate trials, Health-Consciousness is a confounding variable. This shows us that we must probe deeper than surface observations, in order to learn about the hidden causes. We also learn that numbers can lie. If there was no control group, it would be very tempting to conclude that clofibrate really does work, provided that it is taken regularly. It is only the presence of the control group which shows us that it is the health-consciousness of the Adherers which leads to reduced mortality, and not the clofibrate drug. Commitment to false theories can lead us astray.

Concluding Remarks

So what do we learn from the examination of these real world case studies about the discovery of causes? We learn that causes do not lie on the surface. To discover them requires deeper examination of the real world. No amount of fancy data analysis can substitute for thinking about the hidden underlying mechanisms which generate the data. Observations tell us about associations, but not about causation. Associations provide us with CLUES to the hidden causal mechanisms, but one needs to exercise the deductive powers of Sherlock Holmes to deduce the causes from these clues. In particular, we see that Causation requires knowledge/theories about the UNDERLYING, and UNOBSERVABLE mechanisms by which the world operates.

10C: Do Cigarettes Cause Cancer?

In this lecture, we will study a major real-world application of statistics. Our goal is to take a top-down view, a broad perspective on the process of gathering data, and using it for policy. The first example is that of cigarettes and cancer. In the next lecture, we will study the health effects of sugar.

In the middle of the 20th Century, there was huge and controversial debate regarding the question “Does Smoking Cause Cancer?”. At that time, the data showed that smoking was correlated with cancer. The battle was over the question of whether smoking CAUSES cancer? No one knew for sure. There were two major difficulties that stood in the way of finding the answer

1. A Randomized Controlled Experiment was impossible. We cannot choose two groups and randomly assign some to smoking and others to non-smoking for ten years, to observe development of cancer. But, in ANY observational study, there is always a

danger of confounding factors. For example, what if smokers are systematically different from non-smokers – for example, non-smokers are Health Conscious, while smokers are not. Then the treatment and control cannot be matched on all relevant factors, and we cannot learn the true cause of cancer.

2. Another major difficulty was that this is a NEW type of causation. With the familiar deterministic causes, if X occurs then Y follows. If X does not occur than Y does not occur. However, many smokers never get lung cancer, and also many lung cancer cases emerge without smoking. Smoking is neither necessary nor sufficient for lung cancer. This makes it easy to argue that smoking is CORRELATED with cancer, but the real cause is some hidden and unknown factor Z, which is necessary and sufficient for cancer.

It is useful and important to clarify this issue further, via causal diagrams for the two possibilities under consideration:

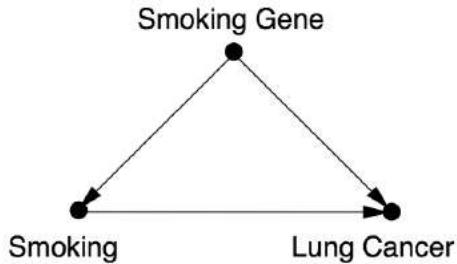


Figure 1: Causation with Confounding

In the above causal diagram, there is a hidden confounding factor which inclines a person to smoking, and also pre-disposes him towards cancer. However, smoking is ALSO a causal factor. Non-smokers have a certain probability of getting cancer, which is higher for those with the smoking gene and lower for those without it, but it is not zero for anyone. Similarly, Smoking increases the probability of getting cancer, and hence is a cause of cancer. The other possibility is that there is only correlation between smoking and cancer, but NO causal relationship. This is represented in the following path diagram

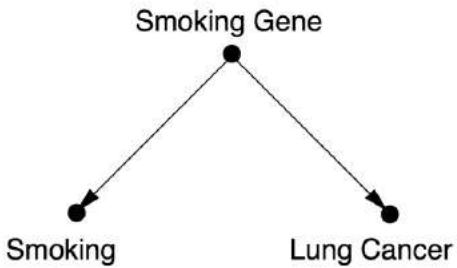


Figure 2 Correlation without Causation

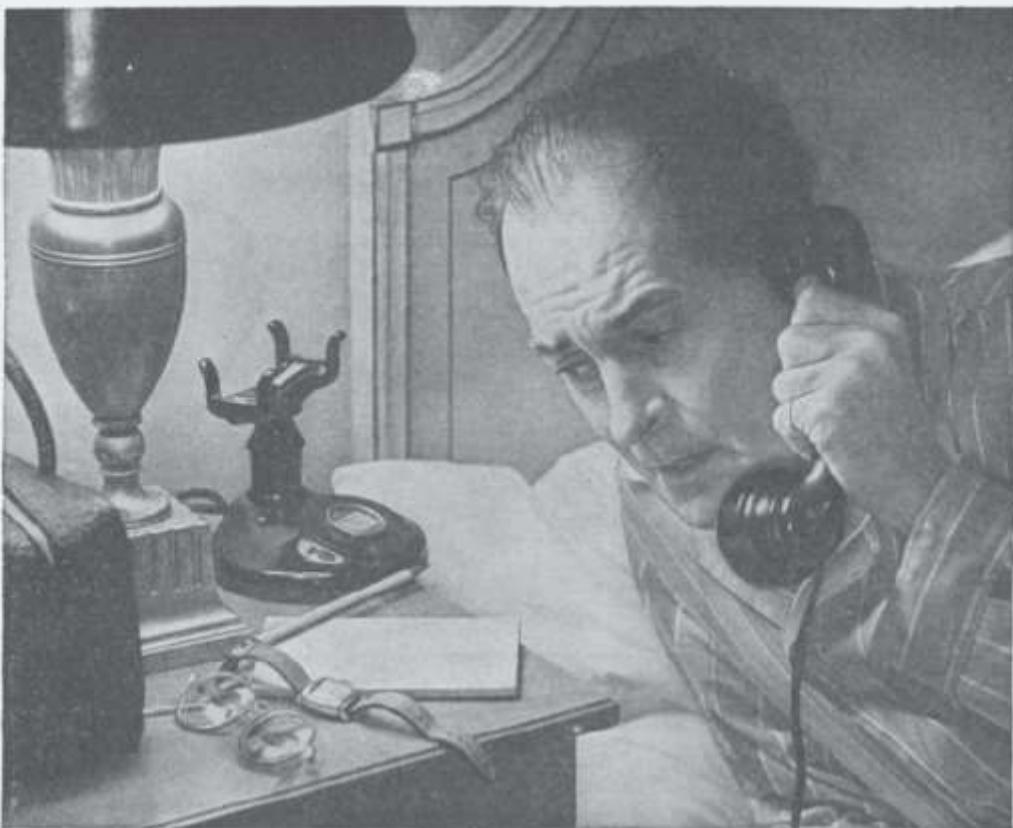
Here the “smoking gene” increases the probability of cancer, and also increases the probability of smoking. Smoking is an imperfect observable indicator of the presence of the Smoking Gene, but does not cause cancer. Only the Smoking Gene is the cause of cancer.

Of course, there is a third possibility that there is no confounding factor at all, and smoking is directly a cause of cancer. Big money was involved in determining the true cause of cancer, and the Tobacco Industry was determined to fight against arguments attempting to prove that Cigarettes cause Cancer. It is very hard to argue for or against the existence of hidden causes, but eventually they were ruled out and it was established that cigarettes do cause cancer. We will cover the history of this debate, based on Chapter 5 of the Book of Why by Pearl and Mackenzie. The goal is to understand how statistics impacts on our lives, within the context of a real-world example.

Background and History of the Debate

In 1902, Cigarettes were only 2% of Tobacco Market, but by 1952 they had captured 84% of the vastly expanded market for tobacco. Two of the major factors in this spread were Automation & Advertising, together with the addictive characteristics of nicotine, which was artificially enhanced in cigarettes for the purpose of creating long-term consumers. Automation was responsible for creating low cost cigarettes for the mass market. Advertising was responsible for vastly spreading the use of cigarettes in the general population.

Impact of Advertising: Nephew of Sigmund Freud, and widely acknowledged father of propaganda, [Edward Bernays applied psychoanalytical insights](#) to get people to buy products without realizing that their minds were being manipulated. In an era in which smoking was considered taboo for women, he created a campaign which equated cigarettes with freedom and liberation, and successfully created a huge new market. Subtle and sophisticated psychological campaigns were used to re-assure the public about the safety of cigarettes. See, for example, the following advertisement from a collection by *Stanford Research into Impact of Advertising*:



"I'll Be Right Over!"

*...24 hours a day your doctor
is "on duty"... guarding
health... protecting and
prolonging life...*

• Plays... novels... motion pictures... have been written about the "man in white." But in his daily routine he lives more drama, and displays more devotion to the oath he has taken, than the most imaginative mind could ever invent. And he asks no special credit. When there's a job to do, he does it. A few winks of sleep... a few puffs of a cigarette... and he's back at that job again...



According to a
recent independent
nationwide survey:

**More Doctors
Smoke Camels
than any other cigarette**

R. J. REYNOLDS TOBACCO COMPANY, WINSTON-SALEM, N.C.

Figure 3 Deceptive Advertising to Reassure Public about Safety of Cigarettes

Over this period of time the incidence of Lung Cancer, once a very rare disease, quadrupled in frequency from 1900 to 1950. By 1960, it became the most common form of cancer. Many doctors began to wonder “WHY did this change happen? What seems obvious to us in retrospect, was not at all clear at the time. British epidemiologist Richard Doll, who eventually discovered the cause, said in 1991, “Motor cars... were a new factor and if I had had to put money on anything at the time, I should have put it on motor exhausts or possibly the tarring of roads.” He teamed up with Dr Hill, famous for his randomized controlled study of streptomycin, to do research on the causes of Lung Cancer.

It was clear that randomized controlled studies were not possible. So, the initial study by Doll & Hill took the form of a “case-control” study. They studied lung cancer patients at 20 London Hospitals. For each person in this “treatment group”, they found another patient matched on age, sex, and social class to serve as a control, but without lung cancer. They then interviewed the patients regarding their smoking habits. Because cancers take years to develop, the issue of how to classify patients as being “smokers” or not required consideration. Doll & Hill decided that anyone who had smoked one cigarette a day for one year would be considered as a smoker. This determination was done by a detailed interview of patients, which asked them about their smoking habits. It turns out that among the 647 cancer patients, only 2 had been non-smokers (0.3%). Among the 647 controls, there were 27 non-smokers (4.2%). We can do the probability calculations to ask if such a large difference can arise purely by chance, under the assumption that smoking has no effect on cancer. The p-value calculated for this event turns out to be 1 over 1.5 million. The relation between smoking and cancer was much stronger for heavy smokers. Thus, the data strongly suggest that smoking is significant factor which is associated with cancer. However, the question of whether cigarettes cause cancer remains open because of the possibility of confounding factors. There are a number of problems with the Doll & Hill study which we now discuss.

Problems with the *Retrospective* Doll-Hill study

In addition to the evidence cited above, the Doll-Hill study showed that heavy smokers had much higher rates of lung cancer than light smokers. Nonetheless, there were many weaknesses in this study which made it possible for the Tobacco lobby to attack it, and question the causal link. One weakness is that it was a *Retrospective* study. This studies current outcomes based on past behaviors. However past behaviors are obtained by questionnaires and interview, and are subject to recall bias. If cancer patients remember or emphasize their smoking history, while the others ignore or minimize it, the survey will create biased results. This problem can be avoided by use of a *Prospective* study, which takes an initial matched group of treatment and controls, and follows them for several years to study their smoking habits and lung cancer outcomes. Another problem with the Doll and Hill study was *Selection Bias*. The sample of patients in hospitals, both treatments and controls, could not be considered as being representative of the general population at large. A third problem was that the study gives us the

inverse probability of the one we want to know. We can learn the probability of Smoking, given that you have lung cancer. But what we want is the probability of getting lung cancer, given that you smoke. There is no way to arrive at an estimate of this second probability from the retrospective Doll-Hill study of Hospital patients with and without lung cancer.

Because of these weaknesses, the strong Tobacco lobby, backed by some powerful statisticians like Ronald Fisher, dismissed all of this evidence. They were secure in the knowledge that causality cannot be proven from observations alone. Even though there were many replications of the Doll-Hill study, biases of the methodology cannot be overcome by replication. The important point to learn here is the interaction between power and knowledge. Changes in policy always affect different groups in different ways, some favorable and some adversely. Reactions to policies are not based on the data alone, but also on interests and power, which allows shaping the discourse. Instead of accepting the statistical evidence for the harmful effects of tobacco, the Tobacco lobby shifted the burden of proof. Stronger evidence for harmful effects of tobacco was required.

The Prospective Doll-Hill Study

Aware of these difficulties, Doll and Hill designed a second study which provided much stronger, almost conclusive proof of the relationship between smoking and cancer. The British Doctors Study was started in 1951, when Doll and Hill sent a questionnaire about smoking habits to all the doctors resident in the UK through the British Medical Register. 59600 men and women answered, but data from subjects younger than 35 years old and from females were excluded since lung cancer was rare in these cohorts and the number of women were too few. As a consequence, they collected 34440 questionnaires from male doctors (10118 born before 1900, 7477 born in 1900–1909, 9459 born in 1910–1919, and 7385 born in 1920–1930) and started long-term observation of their mortality. The choice of this cohort was brilliant: as follow-up of UK doctors is easy, since they have to update their names on the British Medical Register to continue to work. Moreover, the collection of data on the causes of death could have been easier in this group of subjects with routine access to high-level medical care. This study proved to be a landmark, both in terms of establishing the methodology of prospective studies, and in providing very strong evidence of the causal relationship between smoking and lung cancer.

After the first 10 years of follow-up, Doll and Hill reported 4597 deaths, and described an association between smoking and lung cancer: “*we have found death rates per 1,000 per annum from cancer of the lung of 0.07 in non-smokers, 0.93 in cigarette smokers, and 2.23 in cigarette smokers of 25 or more cigarettes. ... Alternatively, we can say that the death rate of cigarette smokers from cancer of the lung has been thirteen times the rate of non-smokers, and that the death rate of heavy cigarette smokers has been over thirty times the rate of non-smoker.*” In such a large sample, there were many cases of doctors who had stopped smoking. The data not only showed a linear relationship between the quantity of cigarettes smoked and the risk of lung cancer, but it also showed that stopping smoking led to a dramatic reduction in risk. This does not seem to be compatible with the hypothesis of correlation without causation (as in Figure 2 Causal Diagram). If smoking is just a marker for a hidden gene, the true cause of cancer, than varying levels of smoking should not lead to varying rates for cancer. Similarly, quitting

smoking should not lead to declining risk of cancer. The Doll & Hill study thus provided very strong evidence for a causal link between smoking and lung cancer.

Further confirmation of the causal effect was added by a paper by Cornfield which showed that the hypothesis of a hidden confounder could not account for the observed correlations between smoking and cancer. It is worth explaining his argument in greater detail. Under the causal assumptions depicted in Figure 2, the only link between smoking and cancer is created by a hidden confounding variable C. This can be depicted as follows:

Smoking <<=> Confounder (C) ==>> Lung Cancer

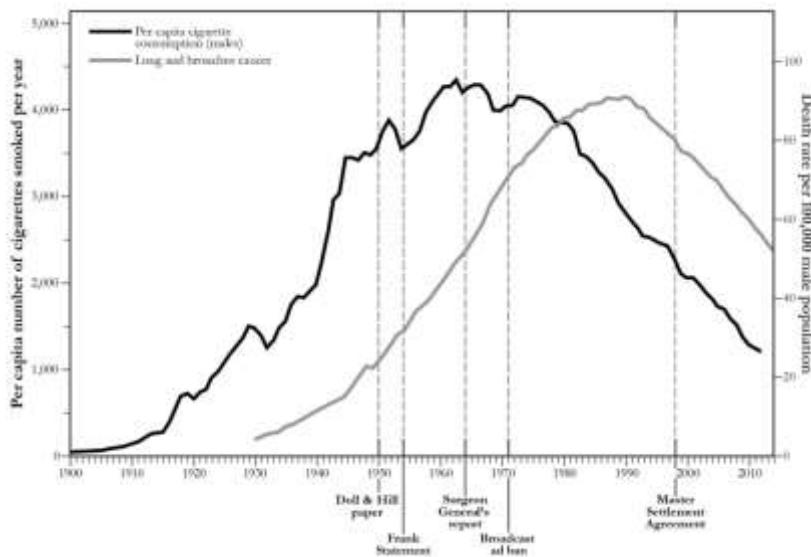
It should be obvious that the observed strength of association between Smoking and Cancer must be weaker than the association between Smoking and C and also between C and Lung Cancer. This is because the overall strength of the chain of association is less than or equal to the weakest link in the chain. If smokers have 9 times the cancer rates of non-smokers, then the hidden confounder must be at least 9 times more common amongst smokers than it is among non-smokers. Neither genetics nor any other “constitutional” factor can differentiate so strongly between smokers and non-smokers. This is because smoking habits are very variable, changing over time and across persons, so that the idea that a fixed constitution corresponds closely to a fixed type of smoking pattern cannot be maintained.

Strong evidence emerging from many different sources convinced those with relevant expertise and knowledge of the statistical evidence of the causal link between cigarettes and cancer. However, the tobacco lobby strongly resisted this conclusion, and put substantial amounts of money to propagate research designed to create doubt about the link. In particular, Sir Ronald Fisher, a heavy smoker, was firmly convinced that there was no causal link. He stubbornly stuck to this position against the rising tide of accumulating evidence. His efforts and the Tobacco lobby succeeded in deluding the public about the strength of the statistical evidence for the causal link. For these reasons, the link between smoking and cancer remained controversial in the public mind long after it had ended among epidemiologists. Even doctors, who should have been more attuned to the science, remained unconvinced: a poll conducted by the American Cancer Society in 1960 showed that only a third of American doctors agreed with the statement that smoking was “a major cause of lung cancer,” and 43 percent of doctors were themselves smokers. This was the first case of organized attempts to deceive the public on a large scale for the sake of corporate profits. Since then, many more campaigns of this kind have taken place – we will cover, sugar, GFC, and climate change later.

The Tobacco Lobby offered the following criticisms of the evidence. They said that we only have statistical evidence; there is no mechanism which has been discovered to explain how cigarette smoking causes lung cancer. However, history offers many examples of discovery of causes by association alone. The disease of scurvy, which afflicts sailors, was discovered by association, and the cure, of eating lemon/limes saved many lives. The fact that this is caused by vitamin C deficiency was discovered much later, in 1930. Furthermore, evidence for mechanisms linking cigarettes to cancer was beginning to emerge in 1950’s. So much so, that even biologists at cigarette companies were convinced of causal link

However, cigarette companies concealed evidence & lied to public. In March 1954, George Weissman, vice president of Philip Morris and Company, said, "If we had any thought or knowledge that in any way we were selling a product harmful to consumers, we would stop business tomorrow." Sixty years later, we are still waiting for Philip Morris to keep that promise. They funded research to show that it was urbanization (and other causes), not cigarettes, which were responsible for steep rise in cancer. Research which came to the "wrong" conclusions was systematically suppressed. They paid heavy consultant fees to statisticians to come up with supporting results.

It is worth noting that we have a stark violation of the famous "Invisible Hand" of Adam Smith. Acting on their personal interests, and using Profit Maximization, the propaganda campaign of Big Tobacco succeeded in delaying action which caused the painful deaths of millions by cancer, while fattening their coffers. Eventually, the evidence became too strong to be resisted. After the Doll-Hill study, and the work of Cornfield and others, the Royal College of Physicians issued a statement regarding the causal link in 1962. The US Surgeon General appointed a commission to study the link and came to the same conclusion. In 1964, an official announcement of the link was made by the USA Surgeon General. This, and other policy measures like ban on advertisements, had a substantial impact on smoking, as shown in the following graph



The tobacco-cancer story continues to this day. It has been estimated that anti-smoking measures have saved 8 million lives over the past fifty years, while 18 million lives have been lost due to cigarettes. So while progress has been made, much work remains to be done. With tobacco fighting every step of the way, the decision that tobacco industry had conspired against the public came only in 2006, opening the way for further punitive measures against the industry. While smoking has gone down in general, but special groups continue to be targeted. In particular, candy flavored like cigarettes has been devised to attract children towards smoking, so as to build life-long consumers.

So, what are the lessons we can learn from this detailed discussion of the history of tobacco and lung cancer? Perhaps the most important lesson here is that it is expensive to gather data and run studies. Anyone who does so has vested interests in the topic. This means that there is no unbiased research, nor can there be. All parties in the debate have interests and biases. The idea of unbiased search through data is a myth. Once we discard the myth of objectivity, we must learn to use statistics carefully. We should not look just at the numbers, but at who produced these numbers, and with what agenda. There are no numbers without agendas.

10D The GNP Illusion

To understand why certain ideas occupy center stage, and others are marginalized, it is essential to study the history of ideas – the “Archaeology of Knowledge” in the terminology of Foucault. When & How did the GDP emerge and become an important tool to guide policy? In the early 20th Century, when Britannia ruled the waves, definitions of prosperity and power involved sea power, coal mines, and other factors which distinguished Great Britain from other countries. The GDP emerged as part of the process of transition of power to the USA, when the two world wars destroyed the economies of Europe. It is partially true that the leaders of the world call the tune, and get to define the criteria which will be used to assess prosperity. However, for our current purposes, another perspective is more useful. Statistics emerge by working BACKWARDS. In the era of British dominance, the question arises as to “WHY does Britannia rule the waves?”. Researchers and economists can see the power of Britain in shaping global affairs, and seek to understand why in quantitative terms. They would like quantify this power, and search for factors which would explain it. These factors then are incorporated into statistics which define power or prosperity. Contrary to our naïve views, statistics are not “objective” numbers which present themselves for our analysis – they are chosen and shaped to quantify intuitions and informal preconceptions about how the world works.

The GDP was adopted at Bretton-Woods 1944, after ruin of European economies. The GDP measures the size of the economy by the total production of goods and services for final consumption (not intermediate goods, used for production of other goods). It is an imperfect indicator of power and prosperity, and measures one aspect of the economy, while neglecting many others. As long as we understand its limitations, GNP can be a very useful number to know. Problems arise when we confuse indicators with reality. YET this is done all too often, with disastrous results.

It is useful to name this phenomenon of substituting the statistic for reality as “The Pygmalion Effect.” In Greek mythology, Pygmalion was a Cypriot sculptor who carved a female statue out of ivory and fell in love with it. After a prayer to the goddess of love, Aphrodite caused the statue to come alive. In a similar fashion, Statistics are created as rough and imperfect guides to complex reality. However, they have a tendency to become substitutes for the reality, with harmful effects. This is another sense in which statistics can be lies – when we think GNP is prosperity, instead of thinking about the number as an imperfect measure of a deeper and more complex concept. We give some examples of the common confusion of Indicators with Reality.

Intelligence is a complex and multi-dimensional attribute, which certainly cannot be reduced to a single number. The IQ Test captures some aspects of intelligence, but misses many others. Similarly, the SAT was devised to assess the capability of students to benefit from a college education. However, the most important factors for academic success are motivation and perseverance, which are qualitative and unmeasurable. In assessing faculty for promotions, recognition, rewards, we would like to assess the quality of their research. Unfortunately, this is an unobservable and qualitative variable. Imperfect indicators like number of publications and academic influence can be found. However, once these indicators are MADE the criteria of quality, then people rush to publish low quality papers in fly-by-night journals, in order to increase the count. The CAUSATION runs from QUALITY to Publication Count and creates a weak correlation. When we strive to increase correlations, then we are trying to reverse the causal arrow, and improve quality by increasing quantity. This does not work. For more discussion, see “[Beyond Numbers and Material Rewards](#)”.

We are taught to believe that statistics are objective, and that we can learn about reality by looking at the facts. This is not true because the number of relevant facts is too large for us to assimilate and absorb. When we start to gather data, we do so in light of some preliminary ideas, pre-conceptions, and theories about what matters and what does not. These theories may be modified in light of the data we gather, and these modifications may lead us to search for new and different kinds of data. Our intuitions and qualitative knowledge of the real-world context play an essential role in determining the kind of data we gather.

In [GDP: A Brief History](#) opens with “Out of the carnage of the Great Depression and World War II rose the idea of gross domestic product, or GDP: the ultimate measure of a country’s overall welfare, a window into an economy’s soul, the statistic to end all statistics. Its use spread rapidly, becoming the defining indicator of the last century. But in today’s globalized world, it’s increasingly apparent that this Nobel-winning metric is too narrow for these troubled economic times.” The GDP measures the size of the economy, as quantified by total purchases of final goods and services on the marketplace. But is this a good indicator of underlying reality we are trying to measure? It has become increasingly apparent that the use of this one indication has led to disastrous results. The GDP conveys useful information about an economy, as long as we understand its limitations and supplement it with other suitable relevant information. Unfortunately, due to the Pygmalion Effect, it became the obsession of planners all over the world. Breaking with this trend, in **1972**, upon being named king of Bhutan, **Jigme Singye Wangchuck** declares his aim is not to increase GDP, but GNH — “gross national happiness.” Despite this solitary objection, policy makers around the globe have sought to maximize growth rates of GNP. Because the GNP is blind to many essential aspects of our human lives, economic policies have not paid sufficient attention to critical problems. Some of these defects of the GNP are discussed below.

Exploring the link between happiness and GDP, Richard Easterlin discovered the “Easterlin Paradox”: there have been massive Increases in consumption of goods and services (GDP) but no corresponding increase in happiness. Similarly, across countries, there are widely varying levels of GDP but these are not reflected in levels of happiness. There is no long-run

correlation between GDP and happiness. This comes as a shocking surprise to economists – all our efforts to increase GDP per capita are worth nothing! We have been wasting time and energy on a futile number chase. Subsequent research on the important question of “Why?” there is no correlation came up with several explanations.

1. Setpoint Effect: As standards of living increase, we get accustomed to them. Happiness results from increases above this new normal, while going below leads to unhappiness.
2. Relativity: We compare our consumption levels with others around us. Because of this relativity, increasing standards of living create no additional happiness on the average.
3. Long-run versus Short-run: Material goods and comforts bring short-run happiness. In the long run, happiness depends on social networks and certain character traits like gratitude.

One of the Important reasons that economic growth has not brought the sought for benefits is because the distribution of the fruits of growth has been highly unequal. Those who have get increasingly riches, while some portion trickles down to the bottom 90%. It has been possible for a long time to feed, clothe, house, educate, and take care of the health needs of the entire planet. This has often been given as the justification for economic growth. The only way to achieve this goal is re-distribution – take from the wealthy and give to the poor. However, economic theory is strongly opposed to this idea. One of the reasons that economists have been able to resist demands for equity is that the GDP statistic is blind to inequality. The GNP is measured by aggregating all the wealth, so that distribution is removed from the picture. Also the GNP aggregates luxuries and essentials all at their market values. So \$1000 worth of bread to feed the poor counts for much less than a single \$10,000 luxury briefcase for the rich. So when we attempt to maximize growth of the GNP, we are automatically driven towards strategies for enrichment of the rich.

Recognizing this bias of the GNP, Mahbubul Haq created an alternative known as the Human Development Index (HDI). This adds measurements for health and education to the standard GNP. This is far from a satisfactory treatment of the problem, but it was a pragmatic compromise to gain acceptability. This move was highly successful, and resulted in greater attention to health and education, factors of great importance for the poor. Later, Amartya Sen went on to develop the capabilities approach, which was a radically different vision for development. Instead of paying attention to accumulation of wealth, he argued that we should look at the development of human capabilities enabled by the economic system. This builds on the Human Development foundations created by Mahbubul Haq, and is strongly aligned with Islamic views on development.

The fact that the global pursuit of GNP growth has led to great harm has been recognized by many. The president of France appointed a commission to study the defects in the measure and propose alternatives. This research has led to the book “Mis-measuring our Lives” by Stiglitz, Sen, and Fitoussi. This proposes changes in three dimensions. One is corrections of the classic GDP measure to take into account several shortcomings. The second is to focus separately on human well-being, which has many non-economic components. The third

dimension involves sustainability, driven by the ecological collapse caused focus on maximization of present wealth, without regards to the future.

Contrary to what is taught, data itself embodies judgments regarding what is and is not important. Developing alternatives to conventional statistics is an attack on existing power structures. The Tobacco Lobby carried out a disinformation campaign against the statistical evidence linking cigarettes to cancer. For some time, the Seven Sisters of Big Oil have been systematically creating doubt about climate change, since measures to produce sustainable development are likely to involve reduction in oil sales and profits. Adding measures of education, health, and sustainability to the GDP will bring benefits to the poor, but will also cause losses to the rich. Applying statistics to make policies in the real world is inherently and unavoidably political. Real world statisticians should be aware of these dimensions, which go beyond the data, in order to make positive change. Today, the most powerful global movement to go beyond GDP is represented in the Sustainable Development Goals, which have been agreed to globally. Policies aimed at improvements along the 17 goals adopted would affect the lives of billions of poor on the planet. Working on developing statistics which capture these dimensions is essential to improved policy, because these numbers guide the policy makers. One of the most successful applications has been the development of the MPI – the multi-dimensional poverty index – which has been used to improve the lives of the poor along multiple dimensions. But a huge amount of work remains to be done. Dimensions of our lives not covered by the GNP include Peace, Justice, Security, the Biosphere, Flora and Fauna – all of these areas are covered by the 17 SDG's. Instead of accepting whatever data is given to them, and doing analysis confined to the data, statisticians need to understand that data is gathered on basis of agendas and pre-conceptions. Thus, we need to be involved in discussing and clarify goals of data analysis, in order to enable policies which improve the lives of human beings.

Concluding Remarks

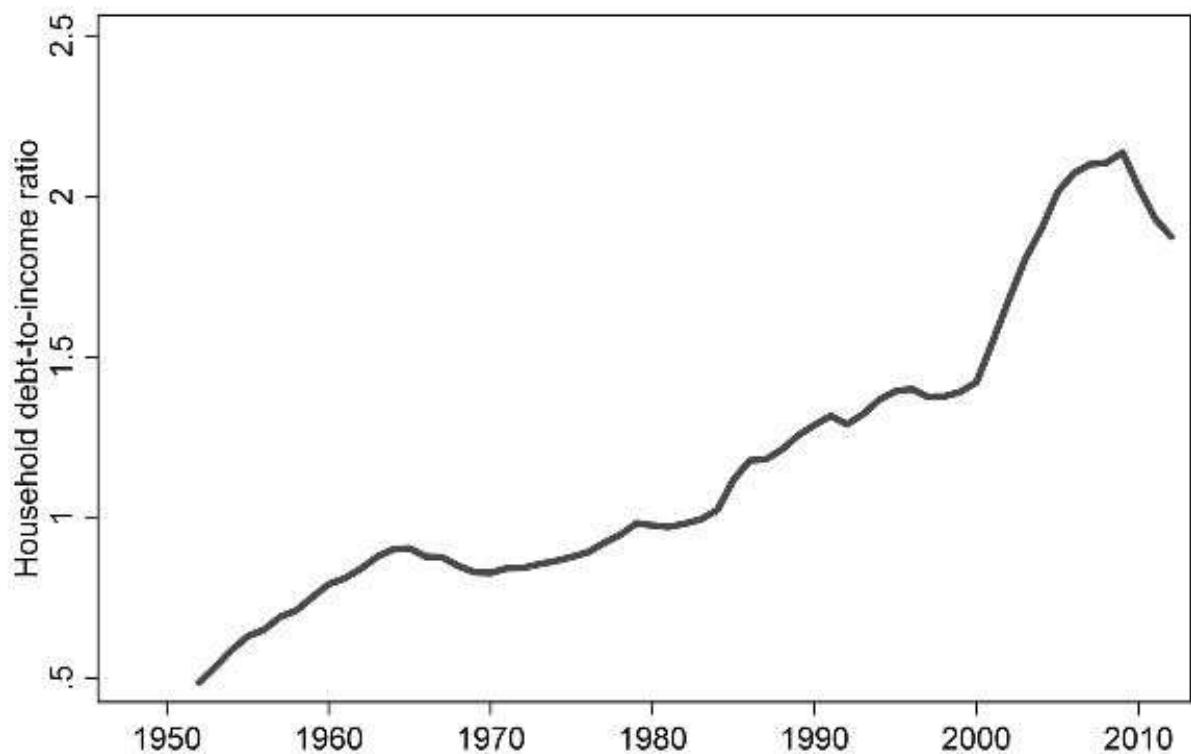
We attempt to summarize the major points made in this lecture. First, what is measured, and what is NOT measured, is all governed by interests. There is not, and cannot be, a “neutral” and “objective” study of the facts. This is because there are billions of facts and it is impossible to study them all. Necessarily, we go in with preconceptions of which facts are important and which are not, in order to extract statistics of value. However, when we look at the data, we may find that it is in conflict with our preconceptions. Then we can modify our ideas and search for new numbers to verify our new hypotheses. Even though the numbers we use – the statistics – are factual, they are selected from among millions of alternatives on the basis of our preconceptions regarding what is important to measure. Everyone who puts in time and effort into gathering facts has an agenda. Statisticians cannot be blind to agendas. Understanding numbers requires going beyond numbers to the discourse in which these numbers are being used. It is essential to understand statistics as a modern form of rhetoric, where numbers are used to support arguments.

10E Causes of Global Financial Crisis: House of Debt

Lec 8E of Real Statistics: An Islamic Approach discusses the causes of the Global Financial Crisis 2007, covering the first three chapters of The House of Debt (HoD) by Atif Mian & Amir Sufi

The first chapter of HoD opens with a MYSTERY. It describes the human misery caused by loss of 20,000 Jobs in a small town in Indiana. On the macro level, 8 million jobs were lost, and 4 million homes were foreclosed. Hunger & Homelessness in the USA reached the highest levels seen since WW2. Why? There was no apparent cause for this economic disaster; that is, there was no war, revolution, earthquake, tsunami, or pandemic, which would lead to massive loss of capacity for production. The economy was fully capable of producing and providing jobs to all – as evidenced by the fact that it was doing so for decades prior to the Great Recession. So, what happened that made it unable to do so?

Mian & Sufi start with a quote from Sherlock Holmes: “It is a capital mistake to theorize before you have data”, and proceed to present data. However, the data they present is strongly aligned with the theory they plan to develop. There is so much data that selection of what data to examine must be done on the basis of theories. Our economic theories provide a number of variables of interest, which have been discussed in the literature, and the search for relevant data is always guided by theory. They start data analysis by looking at the dramatic rise in household debt: from \$7 Trillion in 2000 to \$14 Trillion in 2007:



They note this is similar to GD'29 (Great Depression of 1929): debt tripled from 1920 to 1929. Thus we see that the GFC was preceded by huge rise in debt. The data also shows that GFC started with a huge fall in consumer spending, again matching the pattern of the GD'29.

Of course, this could be just a coincidence, an accidental pattern. Therefore, we examine the International Evidence to see if this pattern is commonly seen globally. We find that Mervyn King, ex-Governor of Bank of England writes in an article entitled "Debt Deflation: Theory and Evidence" that the biggest recessions in the 1990s (those of Sweden & UK) were preceded by biggest increases in debt. Similarly, what [Reinhart-Rogoff](#) call the Big Five: Spain in 1977, Norway in 1987, Finland and Sweden in 1991, and Japan in 1992 are all characterized by a similar pattern. The recessions are triggered by collapse in asset prices, leading to banking crisis, and are preceded by large increases in asset prices and debt. This leads to a natural question: are recessions caused by a Banking Crisis or by increasing amounts of Debt? Here an article by Jorda, Schularick, Taylor ([When Credit Bites Back](#)) which analyzes 200+ Recessions in 14 countries from 1870 to 2008 provides a clear answer: *a close relationship has existed between the build-up of credit during an expansion and the severity of the subsequent recession.*

The data show that buildup of private debt goes with severe recessions, but does not establish causality. There are alternative theories which suggest that buildup of debt is a sideshow, not a cause of recession. Prominent among these is the Fundamentals View, also known as the Supply Side view. During the Great Depression, economists held that the output or GNP would be created by the factors of production capital and labor. Furthermore, supply creates its own demand, so both factors would be fully employed to create maximal output. Keynesian theory was based on the observation that there was excess capacity – both capital and labor were unemployed – contrary to this theory. For the GFC, we note that there has been no reduction in capital and labor, the factors of production – no war, earthquake, or pandemic, which led to destruction of capital, or reduction of labor force. However, modern supply-siders have introduced Rational Expectations as a fundamental factor. If people had a strong belief in future prosperity, this would lead to taking debt to invest in housing (or stocks). However, changing beliefs could lead to collapse. The recession would be caused by the changes in beliefs about the future. High debt occurs as side-effect of these beliefs but is not a cause. A variant of the fundamentals view is that of Animal Spirits. The difference here is that expectations are not rational. In either case, NOTHING can be done about the recession.

Policy response to the Great Recession which followed the GFC was built around the Banking View, which states that the problem is weakened financial sector. When banks suffer heavy losses, they are unable to supply the credit needs of economy, and this leads to recession. The solution is to strengthen financial sector, by bailing out the banks, and making money available to them at low interest rates. This should get the flow of credit going, preventing the recession. Here the problem is seen not as being too much debt, but the opposite. This corresponds to Milton Friedman's views that the Great Depression was caused by a contraction in the money supply created when the FED allowed banks to collapse. If we bailout the banks and restore the money supply to normal levels, we would have staved off the Great Depression.

According to the banking view, there is no such thing as too much debt. The solution to the GFC lies in creating even more debt.

Three major points-of-view emerged regarding the causes of the Great Depression:

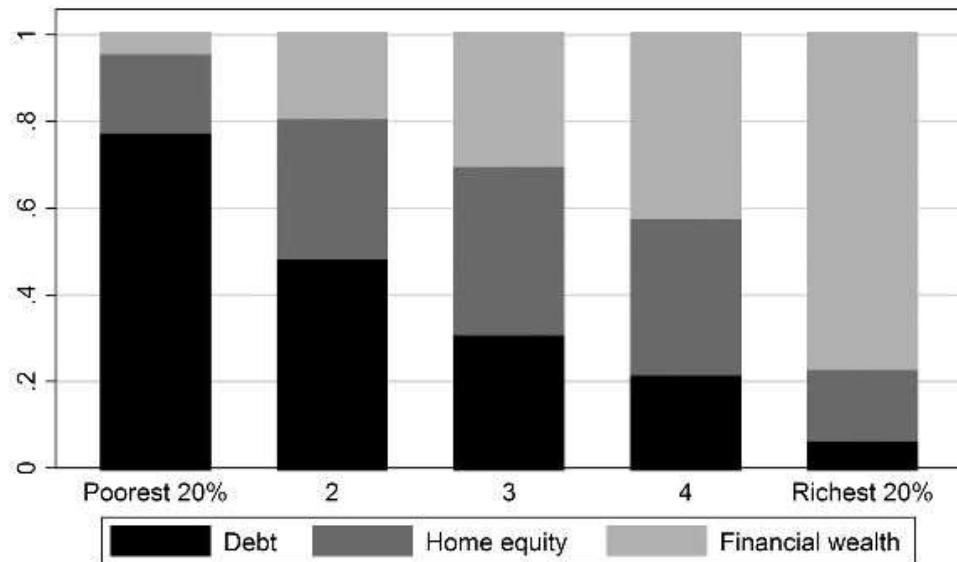
1. Friedman: Monetary Contraction created when FED allowed banks to collapse.
2. Keynes: Collapse of Aggregate Demand (large reduction in consumer spending)
3. Fisher: Debt-Deflation: fall of asset prices leads to increase in real-value of debt.
Consumers save to pay debt, and reduce demand for goods. This leads to further fall in asset prices.

To this day, there is no clarity on which is the correct explanation. The responses to the GFC were crafted by Ben Bernanke, chairman of the FED, and a devotee of Milton Friedman. This consisted of bailing out the banks and making massive amounts of money available to them at nearly zero interest rates (an unorthodox policy labeled Quantitative Easing). Atif Mian and Amir Sufi argue that we have massive amounts of data available to us today, and this makes it possible to empirically evaluate the different hypotheses regarding the causes of the Global Financial Crisis.

The Big Picture which emerges from this investigation is that Interest-based Debt is at the heart of the problem. Debt forces the weaker party to bear risk. Negative outcomes can lead to massive losses of wealth to the poor, leading to falls in consumption and consequent recession. The identification of the cause as the nature of the interest-based debt contract leads to a radically different solution. Instead of bailing out bankers, we need to bail out the borrowers. In addition, it is possible to change from interest-based system to Islamic-style musharaka contracts, which share risk equitably. Such changes would insulate the system from crises. The GD and GFC are man-made disasters, not inevitable. Keynes correctly focused on collapse of aggregate demand, while Fisher picked up on the importance of debt (missed by Keynes). Also, Friedman's theory regarding purely monetary causes, was wrong. We now proceed to do the data-analysis needed to prove these assertions.

We start by noting the harshness and injustice of interest-based debt, as embodied in the home mortgage. A homeowner puts lifesavings of \$20,000 as a down payment to purchase home of \$100,000. The bank provides him a loan of \$80,000 and gets the home as collateral. The mortgage contract guarantees the return to the bank. If housing price declines by 10% to \$90,000, then the equity of the homeowner declines to \$10,000, wiping out 50% of his net worth. If it declines by 20%, he loses everything. If the house price declines by 30% to \$70,000, the homeowner has negative equity. That is, he has to pay off a loan of \$80,000 to purchase an asset worth only \$70,000. Compare this to the Musharaka, an Islamic contract, which makes the bank and the buyer co-owners of the house with equity shares of 80:20. Losses are shared in proportion to the equity share. If the house price declines by 20%, the bank loses 16,000 while the homebuyer loses 4000, in proportion with their equity. Given that borrowers are poorer and lenders are richer, this is a much more just distribution of losses.

We now show empirically how the fall in housing prices wiped out the wealth of the poorest homeowners. The graph below shows the net wealth and debt of households, grouped by quantiles, prior to the GFC:

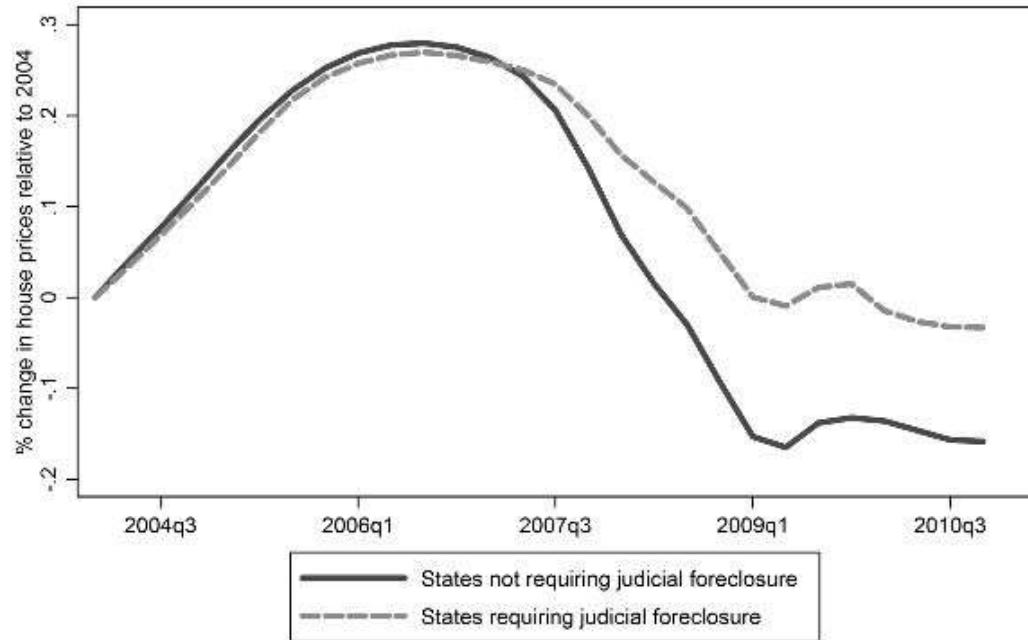


It is clear from the graph that for the Poorest 20%, Debt >> Assets, while for the richest 20% Assets>> Debt. This just reflects the fact that the rich lend to the poor, which is natural. However, the Interest-based Debt, backed by collateral, insulates the rich from risk! This situation is made worse by High Leverage – that is, a high Debt/Assets ratio. This magnifies the effects of small changes in asset prices on the net wealth of the poor. We will now see how these dynamics played out in the Global Financial Crisis, which was initiated by a huge fall in house prices.

From 2006 to 2009, housing prices fell by 30%. This wiped out lifetime savings of the poorest homeowners. Stock prices fell and then recovered. Bonds which offer fixed returns, had increasing prices over this time period, counter-acting the fall in house prices. Of course, Stocks and Bonds are ONLY held by richer households. As a result, net wealth of the richest households was not affected very much by this collapse of housing. Although the average fall in housing was 30%, there was substantial variation across counties. Some had much larger falls, while others were barely affected. These differentials allow us to assess the relative importance of different causal factors in the GFC.

A housing price bubble was created by aggressive lending to high credit-risk households, which were marginal in terms of qualifying for traditional mortgages. This led to higher increases in housing prices in some below median income counties. Correspondingly, there was a larger collapse of housing prices in same counties. On the other hand, some counties were unaffected by these price bubbles. Another factor which caused differentials was foreclosures & fire sales. When homeowners fail to pay mortgages, foreclosure transfers assets from the homeowner to bank. The bank has incentive to sell cheaply and quickly. They have no incentive to maximize profits from the sale, as they only need to recover their equity, and any surplus over

this amount would go to the homeowner. Losses are covered by insurance. But these foreclosure sales have strong negative externalities. Housing prices are estimated by sales of comparables in the neighborhood. A fire sale depresses prices for entire housing market in the neighborhood. Evidence for this effect is demonstrated in the following graph by Mian and Sufi:



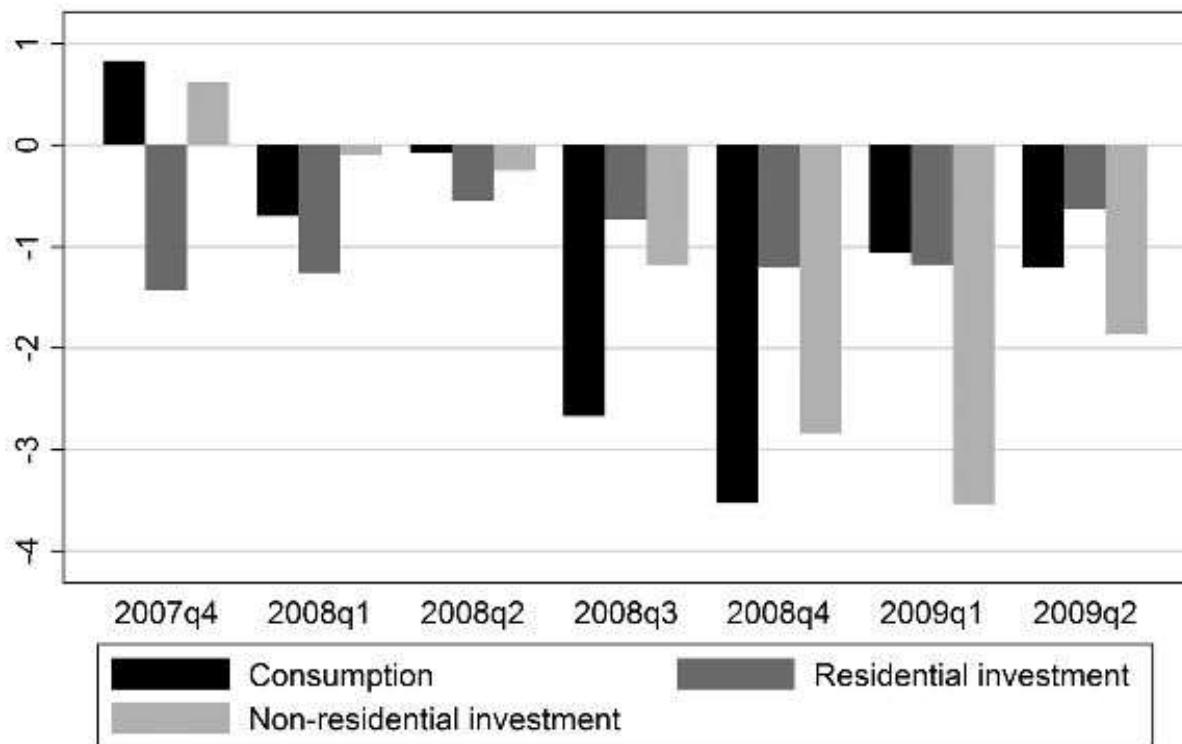
Some states have laws which require court judgment for foreclosures. This is more time-consuming and expensive relative to other states which do not have such requirements. The graph shows that housing prices declined by significantly smaller amounts in states which required judicial foreclosures. This shows how foreclosures have the effect of depressing housing prices.

The central debate addressed in HoD is the Banking View which dominated policy response to the crisis, versus the Levered Debts view which is developed by Mian and Sufi. Note that the banking view suggests bailing out the banks, while the levered debt view suggests bailing out the indebted households as the solution. According to the Banking View, it was the collapse of Lehmann Brothers in September 2008 – rather, the failure of the government to step in and bailout them out – that led to the GFC. This led to a stoppage in flow of credit. In contrast, the Levered Debt view suggests that decline in housing prices wiped out the wealth of the poorer households, and led to substantial reductions in consumption. This was the source of the Recession which followed.

First, the NBER dates the beginning of the recession in Q4 of 2007, 3 quarters before the Lehmann Brothers collapse. Mian and Sufi argue that residential investment should be considered as consumer spending on housing, and this collapsed well before 2008. “*The collapse in residential investment was already in full swing in 2006, a full two years before the collapse of Lehman Brothers. In the second quarter of 2006, residential investment fell by 17 percent on an annualized basis. In every quarter from the second quarter of 2006 through the second quarter*

of 2009, residential investment declined by at least 12 percent, reaching negative 30 percent in the fourth quarter of 2007 and the first quarter of 2008. The decline in residential investment alone knocked off 1.1 percent to 1.4 percent of GDP growth in the last three quarters of 2006.”

The following graph shows the timing of the decline in various components of spending. These timings tend to support the levered debt view, as we now discuss.

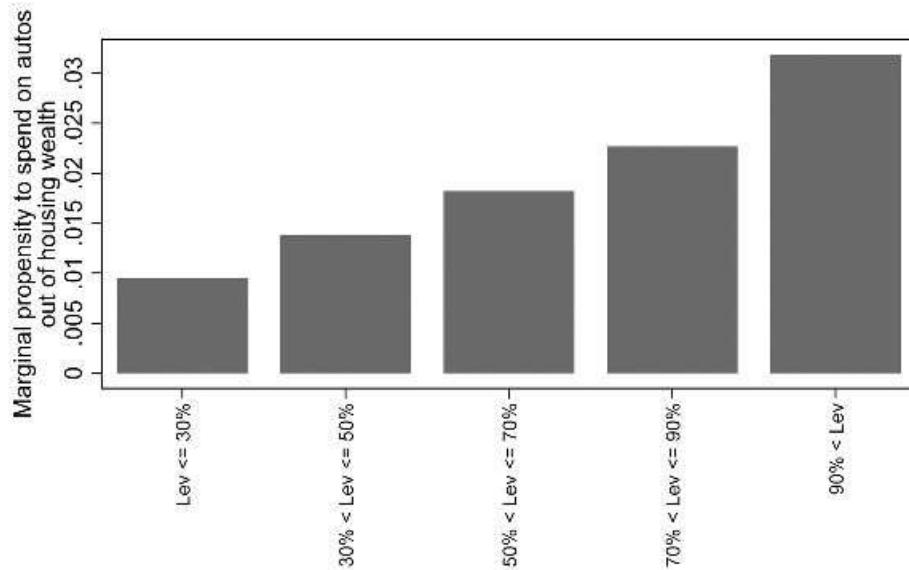


HoD documents declines in consumption spending in different sectors starting from 2007, well before the Lehmann brothers collapse. If the banking view is correct, then stoppage of credit flows to businesses should cause non-residential investment (NRI) to decline. But this started already in Q1, Q2 of 2008, before the Lehmann collapse. It is true that all of the declines – consumption, residential & business investment – became extremely large in Q3 2008, in accordance with the banking view. However, if we look at the pattern of decline in NRI, we see that it follows large decline in consumption. When business saw a steep decline in consumption, they reduced investment, supporting the Levered Debt View. This suggest that it is the consumption collapse which drives the decline in NRI, and not the collapse of credit.

In addition, the geographic patterns (to be discussed) and timings of the decline in consumption and investment are consistent with fall in Aggregate Demand (a la Keynes), and not with fall in banking credit (a la Friedman). While the arguments above are suggestive, they are not conclusive. A **Counter-Argument** favoring the Banking View can be made as follows. The early decline in consumption and investment occurred due to consumer anticipation of banking

crisis. EXPECTATIONS of future collapse played a key role. One weakness of this explanation is that almost no one foresaw the coming crisis (see “[Why did no one see it coming?](#)”). A second counter-argument appears even more powerful, and is harder to handle. The system was sputtering prior to Lehmann collapse, but the biggest decline in spending came in Q3 (July, Aug, Sep) 2008, the same quarter as the collapse. The counter to this is that the banking crisis did start in Q3 and did have an enormous impact on the economy (just like the dotcom crisis of 1999). However, this impact would have been short-lived, and economy would have quickly returned to normal, if not for the factor of huge consumer debt, which led to the prolonged and severe recession which followed. Establishing this requires arguments not covered in the first three chapters of HoD, which are the topic of this lecture.

Conventional economic theory ascribes no role to debt – one man’s debt is another man’s asset, so the net effect should cancel. This is why Irving Fisher’s Debt-Deflation theory for the Great Depression sank without a trace; see [Mervyn King “Debt-Deflation: Theory and Evidence”](#). An important weakness of Keynesian theory is that it misses the significance of debt. Even though debt was an important part of the mechanism leading to severe and prolonged recession, it is only now being re-discovered by some economists. A standard criticism of HoD arguments is that the decline in consumption would come from a wealth effect alone due to collapse of housing prices. There is no relevance of Debt in this argument. There are two important counters to this argument, both of which show that debt matters a lot. The first is the foreclosure externality already mentioned. This occurs due to debt, and substantially lowers housing prices. The second has to do with the distribution of losses in wealth. If the wealth of the poorer households is wiped out, this causes a substantial reduction in consumption. The same reduction in wealth for the richer households would not have much impact. HoD provides some empirical evidence for the impact of this distributional effect in the following diagram, which depicts expenditure on automobile purchases classified by debt-leverage quintiles of the population:



The Highly Leveraged Households have highest Marginal Propensity to Spend on Autos. This means that a fall in net worth of most indebted homeowners leads to highest fall in consumption.

Chapter 3 of HoD concludes with a summary of the Empirical Evidence we have discussed above. Across the globe in general, and for the GFC in particular, severe and long recessions are preceded by a build-up of debt, and are triggered by a collapse in consumption. Asset Price collapse which triggers banking crises has differential effects on different income groups in the population in presence of interest based debt. Debtors are wiped out, while Creditors are protected. Consumption collapses when net worth of large numbers of people in the lower quintiles falls. Consumption is not affected by much when the net worth of the rich falls. This last statement is illustrated by the DotCom crisis of 2000, which saw a loss in value of stocks of roughly equivalent magnitude to the loss of value in housing prices in the GFC. However, stocks are held mainly by the wealthy, and there was no financial crisis or recession as a result.

Concluding Remarks

What are the lessons that we learn from this discussion of the causes of the GFC, which is based on [Chapters 1 to 3 of House of Debt](#)? The first lesson is almost exactly the opposite of the Sherlock Holmes quote used in the opening chapter: “It is a capital mistake to theorize without data (as this creates biases)”. To the contrary, it is impossible to do data analysis without theories to guide us as to which aspects of the data are relevant and important for our analysis. In the modern age, there is too much data to allow us to examine it for patterns without such guidance. Mian and Sufi start by examining the buildup of debt prior to the GFC, but the WDI database lists 1400 data series for 217 economies, and those going for an analysis without a theoretical framework would find many other significant coincidences in the huge data set. As an illustration, no one would think to make a graph of housing price falls in states which require judicial procedures for foreclosures and compare it with states without such requirements, unless we had a theory which tells us that foreclosures lead to negative externalities for housing prices.

These chapters of HoD have been examined in detail because they show how good use of statistics is entangled with theoretical frameworks for understanding the real world. Theories we have about the world guide us to data. In turn, examination of data leads to modifications in theories, which generate new demands for more and different data for evaluation. This in conflict with the idea that we can fruitfully examine data alone, without any knowledge of the real-world mechanisms. We can discover patterns in such analysis, but the meaning and significance of such patterns can only emerge after placing them in their real-world context.

A second important point, which goes against conventional training in statistics, is that data analysis is simple and intuitive. We look at which states had larger falls, who has more debt, etc. Simple comparisons of numbers suffice; no complex theoretical distributions are involved. Conventional analysis substitutes thinking about real world with complex mathematical assumptions about the data, and frequently leads to false conclusions. It is rarely clear to the

researchers themselves how much of their results depend on the statistical assumptions they have made, and how little it depends on the data.

The final point is perhaps the most important. Data analysis is always done within a battlefield of interests. The analysis done in House of Debt speaks for the powerless, the 4 million homeowners who got evicted from their homes. Even though the results are intuitively obvious, this analysis did not make it into the hallways of power. Instead, the self-serving banking view was promoted by the powerful financial lobby. This led to policies which bailed out the bankers by giving them subsidies of trillions of dollars for their criminal activities which defrauded millions of poor. Even today, the perspectives presented above are the minority view, and the banking view dominates. The power/knowledge angle is essential to keep in mind in studying statistics.

Links to Related Materials: Review/Summary of House of Debt:
<https://ssrn.com/abstract=2517476> House of Debt by Mian and Sufi: bit.ly/azHoD

11: Causal Analysis & Path Diagrams

This is the blurb for chapter 11 placeholder

11A: Western Philosophy: Obstacle to Understanding

The goal of the remainder of this course is to explain the concept of causality, and to differentiate it from correlation. Whereas the correlation of X and Y is symmetric, causation is one-directional. Psychologists have found that infants start understanding causality even as early as six months. This is hardly surprising – infants learn that crying loudly will bring an adult caretaker – basic causality. With time, they develop an amazingly sophisticated approach to learning about causal relationships. In contrast, Western philosophers who study the subject have only become increasingly confused. The number of papers written, and the number of positions taken about what causality means, and how we can learn about it, is increasingly large, complicated, and headache-inducing. Why do infants find causality easy to understand, while philosophers have such great difficulty with it? The lecture attempts to explain this mystery.

The story is long and complex. This lecture provides a very brief outline of material which is covered in hundreds of books and thousands of articles. The story starts with the advent of Islam 1450 years ago. The final message of God to mankind promised to give human beings new knowledge. This knowledge, revealed in the Quran, and in the sayings and example of our Prophet SAW, revolutionized the world. It made ignorant and backwards Bedouin thought-leaders of the world. The light of this knowledge eventually reached Europe and ended their dark ages, bringing about the Enlightenment. The flood of knowledge in the books of the Islamic civilization clashed frequently with the teachings of the Catholic Church. The Church attempted to suppress this knowledge but could not succeed. For a diverse set of reasons, Europeans rejected their religion and made science their new religion (see: Corruption of Clergy & Loss of Faith in Europe: <http://bit.ly/azetst2>).

The discoveries of Newton's were tremendously influential in shaping the development of Western philosophy. Rejection of religion led to the rise of materialism. Rejection of God, afterlife, heavens, and the soul eventually led to the rejection of all unobservables. Newton's universal laws of motion showed that every particle obeyed a determinate trajectory. This led the creation of a belief in a deterministic universe. A core commitment to this belief has dramatically affected, for the worse, the development of philosophy in the Western intellectual tradition. The idea that the universe is deterministic conflicts with our strong intuition that we are free to make choices, and the future depends on the choices we make. Nearly all of modern Western philosophy of science is an attempt to reconcile free will with Newton's deterministic universe. Philosophers work on an impossible goal, since one cannot create a philosophy which reconciles truth with falsehood. This crack in the foundations has had profound effects on the entire structure of knowledge created in the West over the past few centuries. Since one cannot use science to derive moral principles, Western intellectuals have abandoned morality on the whole. This has had disastrous effects on the educational process – see the [Marginalization of Morality in Western Education](#). More relevant to our current concerns, philosophical controversies over concepts central to science have continued without satisfactory resolution for centuries. Commitment to determinism makes it impossible to understand the concept of causation, and to differentiate it from correlation. This explains why Western philosophers have not managed to arrive at a satisfactory definition, even though infants understand causality very well. In later lectures, we will explain more clearly and directly how infants understand causality, while philosophers fail.

Slides for this video lecture: [Western Philosophy of Science.pptx](#)

Links to Related Materials:

1. •CK Raju: Is Science Western in Origin: <http://bit.do/azswo>
2. •What the world lost due to the decline of the Islamic Civilization. <http://bit.do/azrdm>
3. •European Transition to Secular Thought: <http://bit.do/etst1a>

11B: Causality as Child's Play

1 The Dilemma of Causality

Study of causality confronts us with a huge dilemma. Intense controversy has raged for centuries over this topic among the philosophers. At the same time, studies of child development show that infants learn about causal concepts almost from birth, and toddlers have a sophisticated approach to causality. How can causality be easily understood by babies, but remain confusing and complicated to the best philosophers for centuries? The difficulty is compounded by the fact that philosophical approaches serve as a basis for empirical data analysis in statistics and econometrics. Even though correct estimation of causal effects is essential for policy, widely used econometric textbooks are deeply defective in their approaches to causality. Angrist and Pischke (2017) examine leading popular econometrics textbooks and conclude that these are based on an outmoded paradigm which ignores causality. They call for a pedagogical

paradigm shift. Chen and Pearl (2013) also examine six leading econometrics textbooks and come to the same conclusion: these textbooks fail to explain central causal concepts with any degree of clarity. Even though Angrist and Pischke agree with Chen and Pearl on the diagnosis, the two sets of authors offer radically different remedies. Since the 1990's Pearl and his group have been arguing for an approach based on Directed Acyclic Graphs (DAGs) as central to understanding causality. Angrist and Pischke (2008, 2013) have written two econometrics textbooks which exposit causality using a "Potential Outcomes" approach, and make no mention of DAGs. Thus, while everyone agrees that causality is very poorly handled in econometrics, there is no agreement about the solution to this problem. This has serious implications since philosophical controversies about causality ramify to the policy context involving real data and applications.

"Inversion" is a favorite philosophical device, where all conventional thinking on a subject is replaced by its diametrical opposite. It is natural to think that adults are wiser than children, because they have the advantage of years of experience and learning. In this article, we propose to look at what children can teach the philosophers – and by extension, statisticians, econometricians, and policy makers. We will examine insights about how children learn about causality from the child development literature, and see how they can be used to clarify philosophical controversies about the topic. This examination gives new meaning to the Biblical "You cannot enter the Kingdom of God unless you became as little Children".

2 What Do Children Know?

The idea that children start out as "tabula rasa" is firmly rejected by child development studies. Babies come into the world equipped with a vast amount of knowledge about the nature of the world into which they have arrived. Their survival depends on their abilities to "root" and "suckle" – to search for the mother's nipple, and to latch on and suck milk. Hespos & Van Marie (2012) review of what infants know and learn about physics concludes that "*The evidence supports the view that certain core principles ... are present as early as we can test for them, and the nature of the underlying representation is best characterized as primitive initial concepts that are elaborated and refined through learning and experience.*" In particular, babies know that there are objects in the world, that these objects persist through time. They can differentiate between solids which retain shape, and liquids which do not. Vision is specially equipped to detect straight lines which often mark object boundaries. Children are able to track trajectories of objects far more rapidly and accurately than experiential learning based on zero knowledge would allow for.

Babies are designed as amazingly efficient learning machines. They have procedural knowledge – what needs to be done to acquire knowledge of the world – which is adapted to the nature of the world they are born into. That is, the procedures they use to acquire knowledge are efficient because of the way that the external world is structured. They also have generalized learning capabilities – this means that even if the world is very different from their inbuilt expectations, they can learn to adapt to radically different environments.

As we will discuss shortly, babies at birth have already learnt to recognize their mom's voice – a task which has only recently come within reach of advanced voice recognition computer programs. It seems fair to argue that the knowledge that went into writing such programs, and something equivalent in computing power to the algorithms they use, is hard-wired into infantile brains. Similarly, babies learn to recognize faces very early. This is another advanced skill, which depends on the knowledge of what face features look like – this is required as cues to learn how to differentiate faces. Such knowledge is built into the facial recognition programs currently in existence, which use AI techniques to derive knowledge from looking at hundreds of thousands of faces. Again, it seems hard to resist the conclusion that some specific kinds of knowledge about the actual world we live in is hard-wired into the infant's brain. For instance, facial expressions for varying emotional states are universal among human beings. As a result, it is possible to hard-wire recognition that a smile represents a happy state. Languages vary greatly but there are general rules which all human languages follow. Babies are born with knowledge of these rules, and learn to parse language into syntactic units, and to differentiate between their native language and other languages with amazing rapidity. The amount of knowledge that goes into programs which can accomplish this is very high. Matching knowledge built into programs which can do what babies do, leads to the conclusion that babies come into the world with an extraordinarily large amount of knowledge.

3 Agency and Causality

Babies learn about causality as soon as they learn to suckle milk. They learn that sucking leads to flow of milk, and stopping leads to cessation of milk. The fact that babies can choose to suckle or not is crucial to learning the causal effect. When they can control the flow, they know that they cause it. If I control an outcome, it occurs when I desire it to occur and does not when I do not cause it to occur. This control allows me to experience my own causal effectiveness as an agent, and is very different from the Humean constant conjunction. A standard objection to this account is the idea that “suckling” is an “instinct” and hence should not count as knowledge. We discuss an experiment which shows that this is not true.

DeCasper and Fifer (1980) conducted experiments on infants as young as 12 hours to learn whether or not they could recognize their mothers' voices. The researchers put a nipple in the baby's mouth which was connected to a voice recording. As long as the babies kept sucking, the voice recording keeps playing, but it stops if they stop for two seconds. After a little while, babies learn that they can control the play of the sound by sucking. Then, when given a choice between their mother's voice and some other, they show a distinct preference for the mother's voice. This suggests that they have learnt the sound in the womb. But for our purposes, the key inference is that the babies learn to cause lengthier play of mother's voice by sucking longer. This is clearly use of knowledge to control an event of type never before encountered, and hence not an instinctive behavior.

One of the reasons that philosophers have difficulty with causality is because they have been persuaded that free will does not exist. Without free will, the baby's choice of whether or not to suckle was pre-determined billions of years ago by the initial conditions at the birth of the universe. Even though this idea is so preposterous that it is not worth taking seriously, it seems to

be widely believed by philosophers, so we provide one more argument for free will. We have direct personal experience of our freedom to make choices, almost from birth. Our lives are built around the choices we make. If this experience is an illusion, then everything we experience is an illusion, and we live programmed lives within a matrix constructed by forces outside our control. We cannot logically rule out this latter possibility. However, there are two types of errors we can make: believing ourselves to be free when in fact we live in a matrix, and the opposite error. The first error is NOT an error because if we believe ourselves to be free within a matrix, then this is also part of what the matrix dictates to us, and we are forced to believe it – we have no choice in the matter. In the second case, if we believe that our actions are pre-determined, when in fact we are free, there is a huge loss to us. We would fail to explore possibilities under the illusion that they do not exist, and that we have no choice. So, in both possible cases, it is best for us to choose to believe in free will.

4 Baby Steps Toward Causality

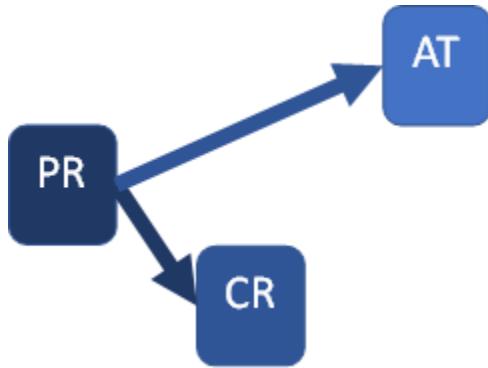
Infants approach the problems of mastering their own capabilities, as well as learning to manipulate the external world, in multiple dimensions. In this section, we present a sketch of how they may learn causality using only “interventions”. This explains how interventions lead to a non-Humean account of causality, and how this is directly tied to counterfactuals. In fact, children have other resources, to be discussed later, which allow them to learn about causality without direct interventions.

Infants are self-aware. They can learn their crying brings a response – an Adult usually appears on the horizon with some uncertain time lag. What is important about this is that the cause is internal to the infant – they are aware that it is “My crying” which causes the appearance of an adult. In the pre-causal stage, the infant feels some discomfort and responds by crying. She observes a response: an adult caretaker appears and attends to her needs. The causal path diagram implied by this description is given below:



In terms of observations, the child observes a sequence of events PR followed by CR followed by AT. The child is assumed to be a passive observer; she observed her pain, and her own response in terms of crying, and then the appearance of an adult caretaker – mommy. The critical dilemma which has stumped philosophers for centuries is that, on the basis of observations, there is no way to determine causality within this sequence of observations. There is an alternative causal path that also explains this same observed temporal sequence:

□

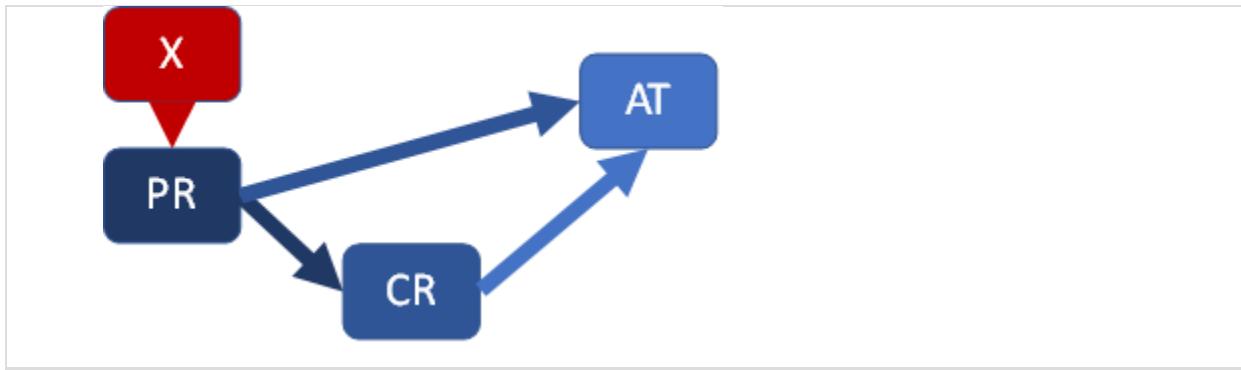


Purely on the basis of observational evidence, there is no way to distinguish between these two possibilities, even though the first possibility leads to a causal link between CR and AA, whereas the second diagram shows a correlation between the two created by the common cause D. In this second causal diagram, CR and AT are correlated because occur due to a common cause PR. However, the infant can go beyond observational evidence by intervening on the CR variable. By doing so, she can easily discover the causal relationship. When the child cries without discomfort – as an intervention, made by choice – the difference between correlation and causation is easily sorted out.



If the causal relationships are as in the first diagram, crying without discomfort will lead to Adult Intervention. In the second diagram, if Adult Attention is caused by discomfort only, and not by crying, then the intervention on crying will not get the desired response. Here we have considered crying without discomfort, which is a common occurrence. The key lesson here is that exogenous interventions are necessary for the discovery of causality. Exogenous means that these interventions are not caused by any of the variables under study. This involves “breaking” the natural causal link between Problem (discomfort) and CRYing, and instead CRYing in response to an exogenous impulse – the desire to summon an adult, or merely curiosity to see the effects of crying.

This leads to a subtle problem in terms of the study of causality. Problem-induced crying may differ from exogenous crying. In fact, this is so common that mothers learn to differentiate between cries based on needs, and crying for attention. Developmental studies establish that children are constantly trying to learn about their environment, and testing different types of interventions to assess the causal consequences of these. Children soon become aware that mommy does not pay as much attention to experimental crying as she does to pain-induced crying. In response to this, children experiment with intervention on “pain” itself – they may voluntarily bump their heads or simulate a fall prior to crying, to see what effect this has:



Intervention in pain – creating it exogenously, instead of natural occurrence – can also lead to clarification of the causal structure of the child’s world. In a complex causal structure like the one above, the child might do multiple interventions. A useful and common intervention is just a variant of the first one: cutting the link between PR and CR. Instead of spontaneously crying without any pain, the child can also suppress crying in response to pain. This is common when the child is engaged in play and gets hurt. Instead of attracting attention, which has the potential of removing the child from the play area, the child may choose to suppress crying, avoiding adult attention, so as to continue playing. In the diagram above, suppressing crying would clarify whether the causal link runs from PR to AT or from CR to AT.

5 Larger Lessons: Counterfactuals & Natural Kinds

The child development literature identifies and studies three mechanisms used by children to learn about causal relations. One of these is “dispositional” – this is the one based on disposition (intentional interventions) by the children themselves, and later, by extension, other agents (humans or animals). In addition, there is good evidence that children can understand fairly sophisticated statistical covariations and use them to assess causality. Children also seem to be born with, or acquire very quickly, knowledge of basic physics – they are surprised when ball A rolls towards ball B, stops short of ball B, and then ball B starts rolling. Contrary to Humean ideas, the causal effects of the balls is understood to be based on contact very early in child development. A survey of the literature by Muentner and Bower argues that the dispositional approach to learning causality is primary, and the other approaches build on this base. For example, Somerville () writes that “Previous work suggests that adults and children readily detect causal structure by intervening on their environment (Gopnik & Schulz, 2004; Gopnik et al., 2004; Kushnir & Gopnik, in press; Lagnado & Sloman, 2004; Sobel & Kushnir, 2003; Steyvers, Tenenbaum, Wagenmakers, & Blum, 2003). Interventions are particularly crucial when one must disambiguate multiple causes or identify variables relevant to causal

outcomes. Critically, interventions enable learners to test causal hypotheses and compare the outcomes of their interventions to expected outcomes (e.g., Sobel & Kushnir, 2003).

Taking self-interventions as the primary source for causal learning among infants leads us to natural answers to two problems which have occupied the attention of philosophers for centuries: Counterfactuals and Natural Kinds. We make brief comments on both.

Counterfactuals: Intervention creates counterfactual knowledge: the infant cries in order to get attention because she knows that she will not get attention if she does not cry. Children are able to use both branches of the causal fork to their advantage – they cry when they want attention, and do not cry when they do not want attention. Because both options can be chosen at will, they are aware of the counterfactual that if I start crying, then an adult will come to see what is wrong, even when they choose not to exercise this option. The children's counterfactual knowledge is far simpler than the philosophers approach. Children's understanding of counterfactuals has been tested directly. In one experiment, subject children were shown movies of a child walking across the floor with muddy shoes. Subjects were all easily able to understand that the mud on the floor was caused by the shoes. Counterfactual questions like what would have happened if the child had removed his shoes before walking across the floor were easily handled by six year old subjects. Harris's (2000) conclusion is that "counterfactual thinking comes readily to very young children and is deployed in their causal analysis of an outcome"

Natural Kinds: To understand the issue, consider David Hume's analysis of how we can learn that when ball A strikes ball B on the pool table, this will cause ball B to move. Suppose an observer watches professionals playing games and observes a thousand such interactions – what can he learn from these observations? Each interaction would be unique, in terms of positions of balls, velocity of A, angle of strike, and configuration of table and other balls. Learning requires encoding the information by throwing away irrelevant details and capturing only those aspects relevant to the causal interaction. A causal model which captures some notion of similarity of objects, especially in terms of having similar causal properties, would seem to be essential to creating such an encoding. In general, inference from past causal interactions to future would require a notion of natural kinds – objects of the same kind are those which have the same causal properties. There is substantial evidence that infants are born with the knowledge that there are objects in the world, and these objects persist over time – they do not blink into or out of existence. Similarly, it seems that some knowledge of natural kinds, and causal properties of objects, is also built into babies. Thus, when we look at "similar" objects, we expect them to have similar causal properties. This notion helps simplify the search for causal properties, and greatly facilitates learning. Suppose 10 events occur at time 1 and 10 events occur at a later time 2. Then there are 1023 nonempty subsets of the initial ten events which could be the cause of 1023 subsets of the later events. It would be impossible to test and explore the one million possible causal hypotheses which would be possible without any similarity relationship. But similarity and localization serve to substantially narrow the space of possible causal connections, making learning possible. By localization I mean that when two balls interact causally, we assume that only the two balls are involved, and the configuration of the remaining balls does not

matter (very much). Without localization, causal interactions could be too complex to allow for learning.

6 Concluding Remarks

What can babies teach philosophers? The ability to control the environment, and to make choices and decisions which affect our personal well-being begins at birth. Babies look for, and suck, when they are hungry, and not when they are satiated. This is not instinctive behavior – it is knowledge, which is built into babies. Thus, tabula rasa hypothesis does not hold. In particular, studies show that infants look for causes for unexpected events. That is, the knowledge that events occur for a cause is built into us. Also, personal interventions provide a powerful means for learning causal relations. If we intervene in the system to bring about a desired outcome, then we learn about causal effects of our actions. This experience of causality is radically different from observations of constant conjunction, which are never sufficient to deduce causality. Making choices, and choosing to intervene, or to not intervene – to suckle, or to stop – are given to us from birth. This free choice is what enables exogenous interventions, and also leads to knowledge of counterfactuals – the consequence of acting or not acting. This free will and resulting counterfactual knowledge is also an essential part of the moral framework within which we live our lives. Moral choices result, partly, from the knowledge of different consequences which occur as a result of our choices.

The causal relations perceived by the baby change rapidly as the baby grows more capable of interacting with the world. Also, because adults are responsive to growth and changes, the same effect can lead to different causal outcomes as growth and learning occur. This means that the concept of causality as a “necessary” connection is a non-starter for infants. In a rapidly changing world, observation of constant conjunctions as a means of learning is also a non-starter. Personal interventions provide for rapid learning of causal connections based on experience. Here, it is important to note the children's "knowledge" is pragmatic: pushing the button causes the toy to turn on. This is not an eternal truth, just something that seems to work for a while -- until the batteries run out, or the toy breaks. The mechanism by which it works is a complete mystery, and will remain so.

In later lectures, we will see how babies generalize from personal experience to other agents. They attribute intentions and causal efficacy to other “dispositional” agents, and are good at reading intentions from actions. Causal learning based on agent interventions can then be generalized to mechanical interactions based on physical laws. To summarize, the results of watching infants learn about causal relationships essential to their survival and growth leads to a framework which is radically different from that of the philosophers. It resolves many puzzles which have stumped philosophers for centuries.

Finally, it is important to note that the infant can learn to manipulate the world while having very little knowledge of how and why these manipulations are successful. When the infant cries and the mother comes to investigate there are two types of causal mechanisms at work. One type is the mechanical aspect – the strength and type of crying, and the transmission of sound to the mother. The second is the dispositional aspect – how and why the mother

responds to the child. Both of these are known to the child only roughly. Bad theories about the mechanisms and the dispositions are compatible with successful manipulations. This is of obvious significance for the philosophy of science.

7 References

De Pierris, Graciela and Michael Friedman, "Kant and Hume on Causality", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/win2018/entries/kant-hume-causality/>>.

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mastering 'metrics: The path from cause to effect*. Princeton University Press, 2014.

Angrist, Joshua D., and Jörn-Steffen Pischke. "Undergraduate econometrics instruction: through our classes, darkly." *Journal of Economic Perspectives* 31.2 (2017): 125-44.

DeCasper AJ and Fifer WP. 1980. Of human bonding: newborns prefer their mothers' voices. *Science*. 208(4448):1174-6.

Chen, Bryant, and Judea Pearl. "Regression and causation: a critical examination of six econometrics textbooks." *Real-World Economics Review*, Issue 65 (2013): 2-20.

Hacking, Ian (1983) *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge university press.

Hespos, S. J., & vanMarle, K. (2011). Physics for infants: characterizing the origins of knowledge about objects, substances, and number. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(1), 19–27. doi:10.1002/wcs.157

Imbens, Guido W. "Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics." *Journal of Economic Literature* 58.4 (2020): 1129-79.

C. M. Lorkowski (2021) “David Hume: Causation”, Internet Encyclopedia of Philosophy, <https://iep.utm.edu/hume-cau/>

Michotte, A. (1963). *The perception of causality*. New York: Basic Books

Morgan, Stephen L., and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015

Muentener, Paul, and Elizabeth Bonawitz (2017) "The development of causal reasoning" in Waldmann, Michael, ed. *The Oxford handbook of causal reasoning*. Oxford University Press.

Schulz, Laura, Tamar Kushnir, and Alison Gopnik. "Learning from doing: Intervention and causal inference." *Causal learning: Psychology, philosophy, and computation* (2007): 67-85.

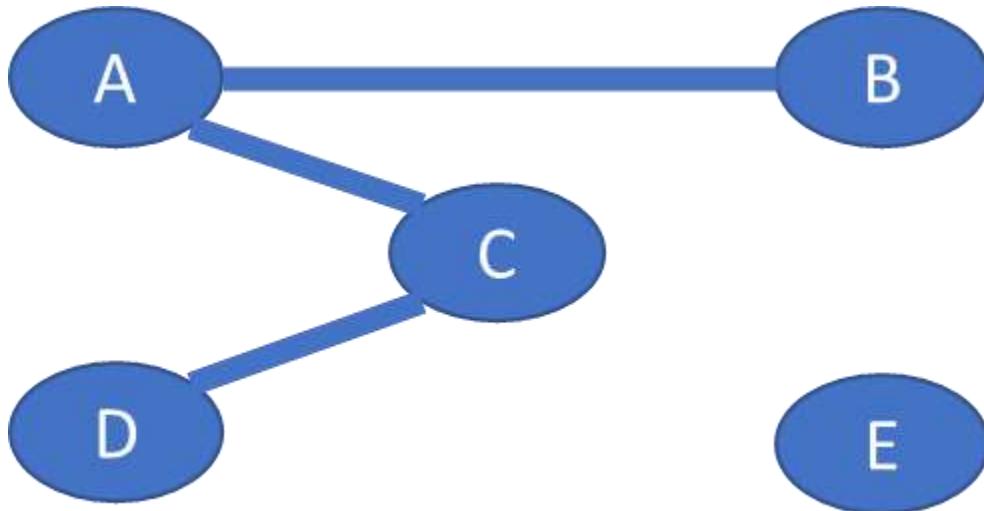
Jessica A. Sommerville, "Detecting Causal Structure: The Role of Interventions in Infants' Understanding of Psychological and Physical Causal Relations"

11C: Causal Path Diagrams

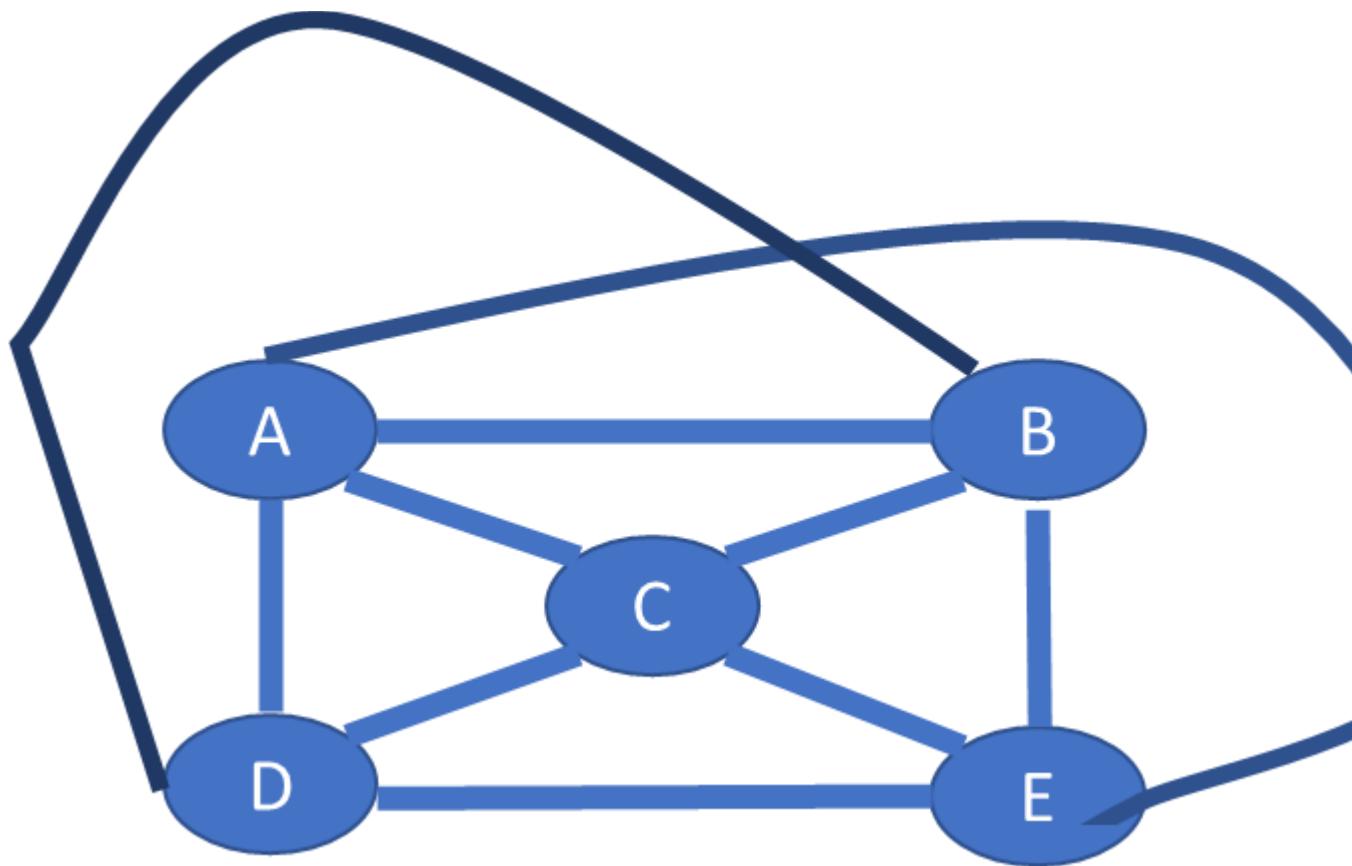
Introduction: In the previous lecture, we used arrows and nodes to represent different possible causal pathways and causal sequences. Judea Pearl has shown that these diagrams are central to the study of causality. This lecture is devoted the development of basic concepts related to causal path diagrams, and how they help us to understand causal relationships.

1 Directed Acyclic Graphs (DAG)

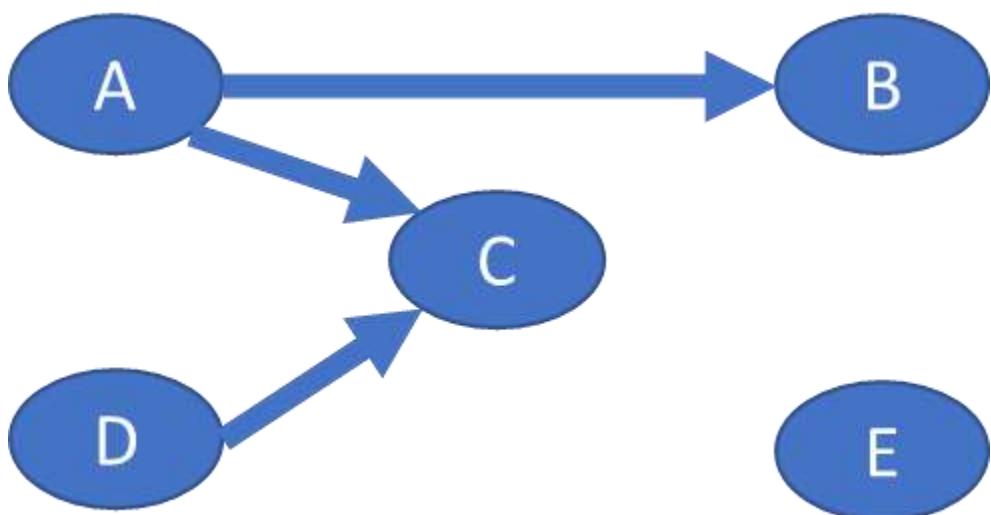
A graph is a collection of nodes and connections between nodes, which are called paths. The nodes will be the variables of interest, and the paths will be the potential causal links. The graph below has five nodes (A,B,C,D,E) and three paths which connect (A,B) (A,C) and (C,D):



The FULL graph has paths between all pairs of nodes; we can draw the full graph for the five nodes above by connecting all pairs of nodes as follows:



Graphs form a natural way to describe complex causal relationships. We will use paths with arrows to indicate the direction of the causal relationship:



This graph describes a situation where A causes both B and C, while A and D both cause C. E is independent of A,B,C,D, since it is not connected by any arrows to the other variables. When every path is marked with a directional arrow, we say that the graph is “directed”. Causal path diagrams must satisfy one additional condition: they cannot contain cycles where A causes B, B causes C, and C causes A. This is taken to be axiomatic, in accordance with our intuition about causality. Imposing these conditions leads to the concept of DAG – Directed Acyclic Graph – which is of fundamental importance in the study of causality. We will use the term “Causal Path Diagram” as a synonym for a DAG.

2 Representation of Causal Relationships

Philosopher David Hume argued that we learn about causation by watching conjunctions – we see event A followed by event B many times, and from this we deduce a causal connection between A and B. This analysis misses, by miles, the central problem. Every event is unique: let event A be billiard ball (8) strikes (9), and let event B be (8) stops moving and (9) starts moving. The fact that event B follows event A does not advance us one tittle or one jot towards understanding causality. A detailed description of events A and B would ensure that we would never see a repetition of these events. For example, if we put colors on the balls, and describe the exact state of wear and tear of the two balls along with the felt on the table, or the exact positions of the molecules within the balls, then event A would be a unique one-time event with no possible repeats. We must use the right type of abstraction to encode the event into a “type” which can be repeated. This encoding must include the causal elements and exclude the non-causal factors. So the color of the ball cannot be included in the encoding, and we can exclude the exact positions on the table of the two balls, taking into account only the mass and the velocity. Including the exact positions of both balls into a description of the events A and B would create great difficulties for us in acquiring sufficient constant conjunctions to learn about the causal connection. So we conclude that:

To discover causal relationships, a first step is to encode every special and particular event into a general event-type which may be repeated. This encoding includes potential causal factors and excludes irrelevant factors, and therefore already embodies prior causal information. Child development studies show that babies learn physics of moving balls very rapidly. Almost surely, a certain amount of knowledge of basic laws of physics governing moving objects and their interactions are built into them. Separating what children know in advance, from what they learn, is difficult, and not necessary for our purposes. Children can learn that there exist objects which retain their physical shape by watching the movements of their own milk bottles, or other constantly present objects. The key knowledge which children must be born with is that similar objects have similar causal properties, and both objects and their causal properties persist through time. Thus, observing a causal effect at one time allows us to predict the same causal effect at a later time. It is not constant conjunctions but an in-built knowledge of “similarity” relationships that enables this causal inference. Observing a causal effect at one point in time, and seeing a similar object at a later time, we can immediately conjecture that the object has the same causal properties. Observation begins with personal experience of causal efficacy of our own actions. Later, a conjectured similarity between ourselves and other agents (human and

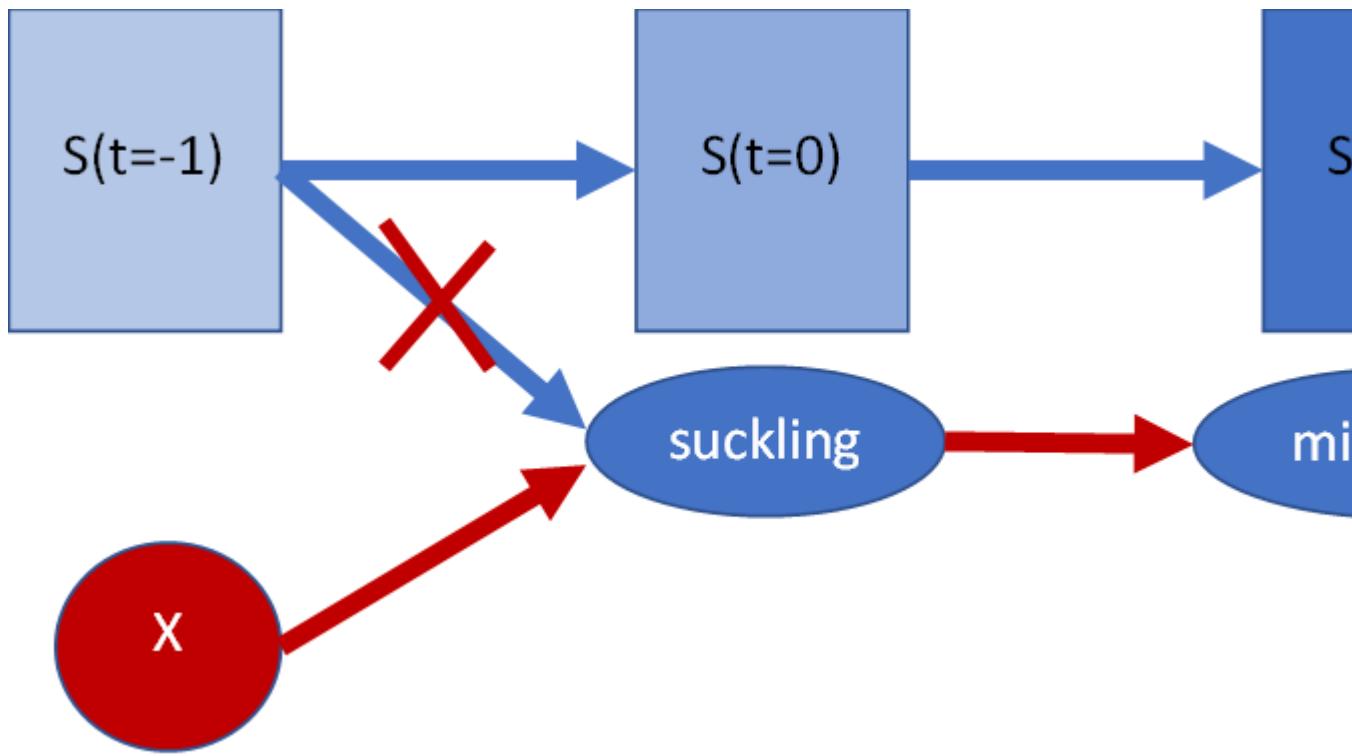
animal), allows us to guess at causal effectiveness of their actions, and also infer their goals from their actions. At the third stage, agents use mechanical objects to achieve remote goals, so causal interactions between objects, initiated by dispositional agents, can also be learnt.

The details of how children learn causal effects are not relevant to our current pedagogical goals. The key lesson we want to extract is that children encode causal relationships into an abstract form – this is the only possible way of learning from specific causal interactions. It seems extremely plausible, and has been argued by many authors, that this encoding takes the form of a causal path diagram, or a DAG. As we have already discussed, a DAG is a collection of nodes (variables under study) connected by causal paths, subject to the constraint that there can be no cycles. There are two fundamental building blocks of a DAG. One block is the node, which built on a concept of similarity – this allows one abstract event to be similar to several specific events and permits inference of common causal structure. The second is the concept of a “direct causal relationship” – this is a single causal path which links two nodes, represented symbolically as $X \Rightarrow Y$ (X causes Y). The entire causal path diagram is built of a collection of such direct causal relationships. We analyze this key relationship further.

3 Direct Causal Relationships

Given a collection of variables under observation (U, V, W, X, Y, Z, \dots), the task of determining the causal structure consists of assessing, for each pair of variable X and Y , whether or not there is a direct causal link between the two. In this section, we examine the question of how we can determine if X “directly” causes Y . We will clarify the meaning of “direct” causation, and distinguish it from indirect causation in the discussion below.

As discussed in previous lecture, infants learn about causality by interventions. They suck and milk flows, so they learn that sucking causes milk to flow. It is our ability to freely choose whether or not to intervene which allows us to assess the causal affects of the intervention. Since this is a key point of confusion for deterministic scientists and philosophers, it is worth elaborating here. Let us use $S(t=0)$ to denote the state of the universe at time $t=0$. Assume that there are physical laws such that $S(t=1)$, the state of the universe at time $t=1$, is determined by $S(t=0)$. Random events like those posited by quantum mechanics do not affect this argument, so we will ignore them in what follows. $S(t=0)$ includes all events which occurred at time $t=0$, which include, let us say, the decision by the infant to suck. Suppose that $S(t=1)$, the state of the universe at $t=1$ includes the causal response, the flow of milk. In what sense, if any, can we say that milk flow is caused by suckling? There is a set of laws which governs the evolution of the universe, and all events at time $t=0$ collectively cause all events at $t=1$. I cannot see any way to make sense of a localization which says that one small micro event at $t=0$ is the cause of another small micro event at $t=1$.



Exogeneity of the intervention does succeed in defining causation: When suckling is chosen as an act of free will, this breaks its connection with the past state of the universe. In this case, the occurrence of milk flow in response, can be unambiguously attributed to suckling. Also, if stopping suckling stops milk flow and re-starting re-starts it, then we can safely deduce that other states of the universe which co-occur with suckling are not causes of the milk flow. This exogeneity – freely choosing to suckle or not – is crucial in localizing causation and connecting my action with its consequences. We suppose that an act of free will is not determined by the state of the universe and the laws of physics – that is what it means to be “free”. This is because we experience our capability to choose actions, and we experience changes in the environment as consequences of our actions. To override this direct experiential knowledge on the basis of a misunderstanding of evolving laws of physics, where undiscovered particles, forces, and laws abound, is an act of faith far deeper than that which religion demands of us.

Assuming that we encode causal relationships as DAG's, the question is how do we discover whether or not there is a causal path between variables X and Y which we observe. Note that the concept of “variables” already encodes a similarity relationship based on abstracting from specific to causally related objects. Thus the construction of the nodes in a causal path diagram also requires some information about causality – this may be what is referred to in the Quran as the teaching of the names to Adam AS. If we know the names – the categories of causally similar objects – we can create the nodes in the causal path diagram. Assuming that NODES (Event Types) are given to us in advance – built-in knowledge – the next step is to learn how to draw directed arrows representing causal connections between two event types X and Y. There is a progression of learning steps involved.

The simplest method of discovery is personal experience. If I take an action, and it achieves my desired goal, then I know that my action causes the desired result. I experience the efficacy of my action as a means to achieving my goal. Here free will and free choice of goal are crucial. I know that my goals are personal to me, and are not part of the universe out there. I know this because I can observe my interiority, and I can observe the way I form and change my goals, sometimes whimsically. When I randomly change my goal I know this change is not caused by any outside forces. When I take an action to achieve my goal, this action is not predetermined by universal forces because it has been chosen by me to achieve an effect known only to me. This makes the action “exogenous” – uncaused by other external variables. To put it more simply, the baby learns how to make the milk flow by suckling, and make it stop by not sucking. She knows that she herself causes the milk flow by personal experience.

Child Development studies show that babies learn very early to guess at goals of other human beings. If human beings take goal directed actions, and achieve their goals, babies learn about the causal effects of actions in achieving goals. Sophisticated studies show that children pay a lot of attention to goals of others, and the actions they take to try and achieve their goals. Even if they fail to achieve the goal, children might repeat the action, since they associate a causal effect with actions directed toward achievement of goals. When no causal agent is present, then children do not deduce causal effects of one unusual event followed by another – they assume it was accidental.

Initial development of causal learning comes from personal goal directed behavior – “making things happen” – in words of Woodward. This is generalized to observations of actions of other agents which appear to be effective in achieving their goals. Later, they observe sequences of events initiated by a causal agents, and apply causal effectiveness to these intermediate agents. For example, if a human hits one ball with another to cause the second ball to go into a pocket, the children learn that the ball can act as a cause. The goal of the human being is transferred to the ball as an intermediate agent, without any volition on part of the ball. Children distinguish clearly between objects – grasping for them – and agents – cooing at them.

Using these basic building blocks, one can arrive at many higher level inferences about causal relationships. However, since our focus is on using data analysis for policy purposes, goal-oriented actions which may be able to achieve desired outcomes is suitable as our definition of causality. This is very different from most standard definitions of causality found in econometrics, statistics, or philosophy. We will show one example of how this kind of definition can be used to deduce causal relationships in a practical context.

Causal Effects of Monetary Policy

One of the key claims of “Real Statistics” is that we cannot do data analysis without understand the real world structures which generate the data. This means going beyond the numbers to the underlying realities being measured by the numbers. This is especially true for causality. To understand the causal effects of money, we first go through some theories about money and its effects on the economy.

1: Conventional Theories: Quantity Theory of Money

To the central question under investigation: “What effect does Money have on Economy?” QTM answers that “Doubling Money doubles prices, and has no effects on the real economy.” Another way of saying this is that “Money is a veil.” Money and Banks do not matter for the real economy. Because it is so easy to demonstrate real effects of money, a more sophisticated version of the QTM allows for short term and temporary effects of money, but denies any real effects in the long run. This is why Keynes states, early in his revolutionary book on “The General Theory of Employment, Interest, and Money” that money matters, both in the short and the long run. However, this fundamental insight of Keynes was rejected later on, in the 1970’s – see “The Keynesian Revolution and the Monetarist Counter-Revolution” (<http://bit.ly/KRMCR>). This is why Romer says that we have been going backwards, and losing precious knowledge in macroeconomics, in the past few decades. Why are conventional theories of money so absurd, and so dramatically in conflict with facts? ONE part of the answer is the support these theories provide to the interests of the rich and powerful. For a deeper understanding of this aspect see: The Battle for the Control of Money. <https://sites.google.com/site/economicsislamicapproach/l124battle>

A second aspect is that extremely poor understanding of causality, and extremely poor statistical and econometrics techniques make it easy to support absurd causal claims, and prevent the discovery of valid causal relationships.

2: Establishing the Causal Effects of Money

Paul Romer, in “Trouble With Macro” (see: <http://bit.ly/ACRomer>) wants to show how Modern Macroeconomists ignore stark conflicts between theory and observation. He cites leading macroeconomists who claim the money has little or no impact on the economy, based on QTM. THEN, he wants to show that monetary policy has dramatic effects on the real economy. Our concern here is to see HOW he does this – Because the method he uses is very strongly matched to our proposed approach to causality, and VERY DIFFERENT from many different philosophical and statistics/econometrics perspectives.

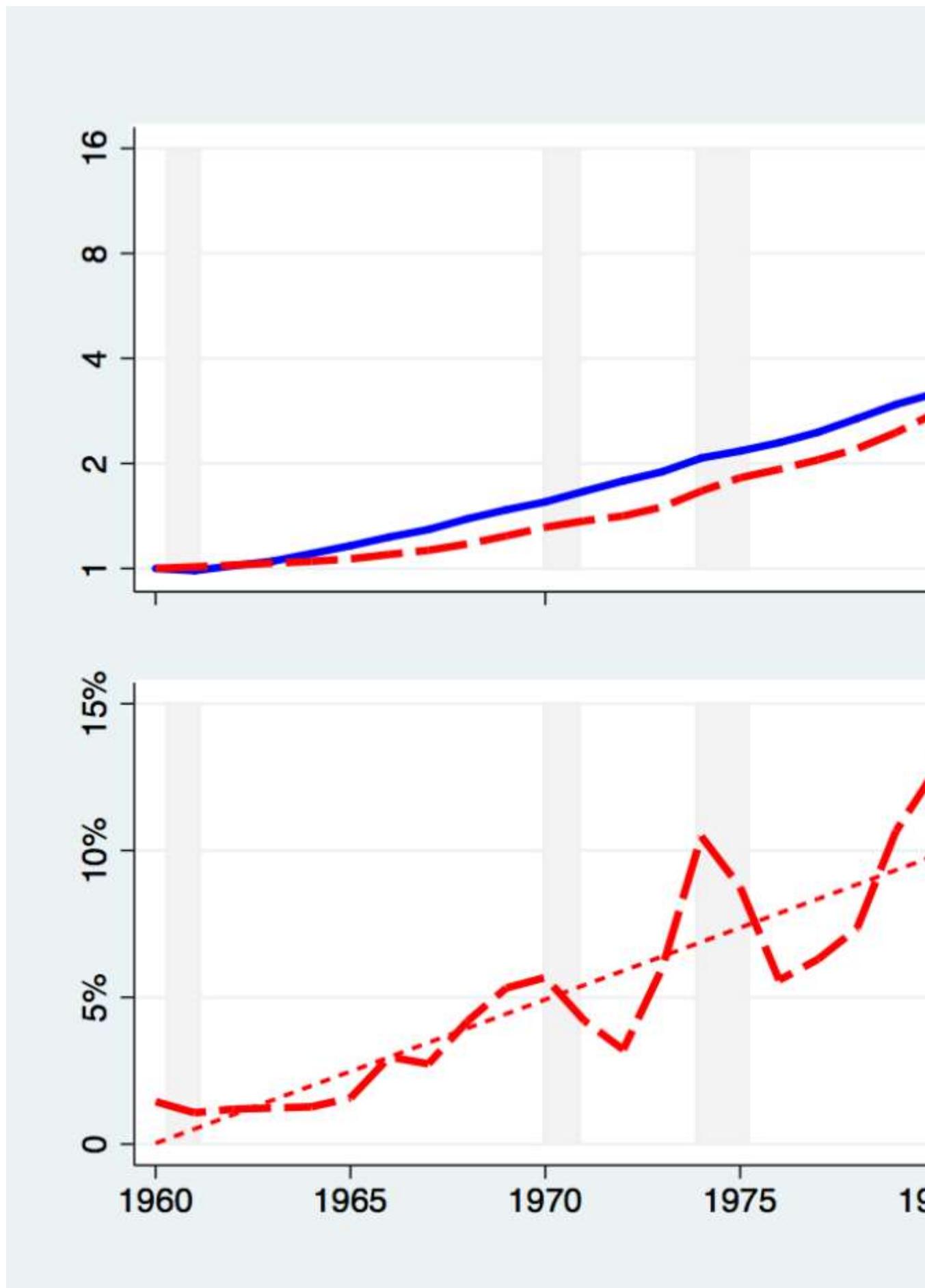
In 1979, when Volcker took charge of the Federal Reserve, he made major changes in the methods of operation. He ANNOUNCED that he planned to increase real interest rate, in order to control inflation. In terms of our analysis of causality, he announced a future goal. He also explained the action he planned to take, to achieve that future goal. As it turns out, the action taken achieved the intended goal. This is very strong PRAGMATIC evidence for the causal effect of the action in achieving the goal. Heuristically:

1. If action achieves goal, a PRAGMATIC causal link between action and goal is established.
2. If action has no effect on target variable, we TENTATIVELY reject causal link.
3. Action can also have contrary-to-expected effect. Causal link exists but works in the wrong direction. OR there are other explanations.

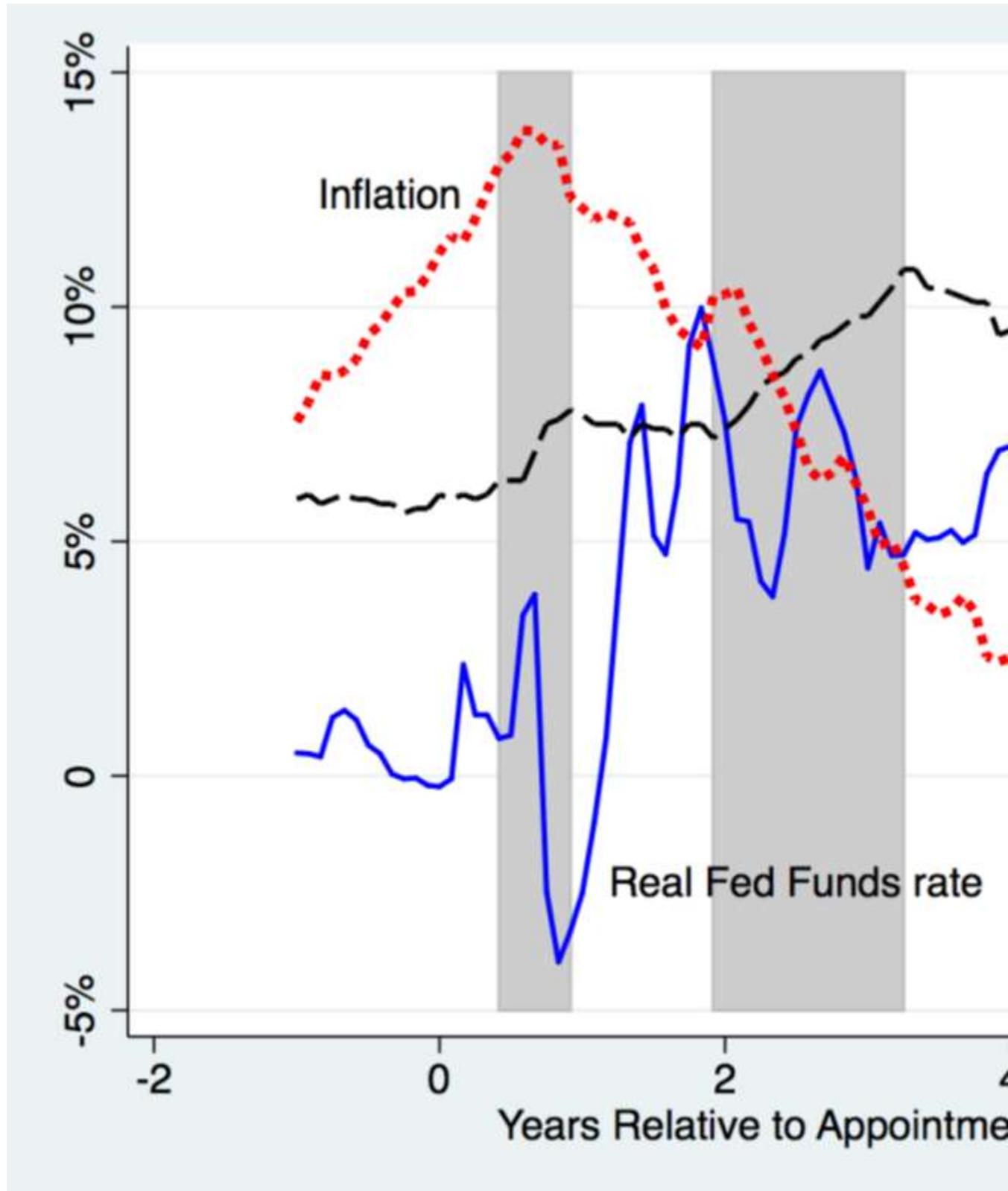
Note that we deduce causality from ONE event! Contrary to Hume, we do not observe repeated sequence of events. If I can control something with my actions, then I know of causal effect. If sucking leads to milk, then I learn that sucking causes milk to flow. I clap my hand, and thunder immediately follows. I conjecture that clapping causes thunder. I do repeated experiments, but soon learn that I cannot control thunder by clapping. To the contrary, I can control the light by flipping the switch. If I can MAKE THINGS HAPPEN, then I learn about causal effects of my actions.

3: What Volcker Did:

The Blue Line is the real Federal Funds Rate, which is SET by the Federal Reserve. The two shaded grey areas are the two Volcker Recessions. The first one was cut short by orthodox Keynesian policy of cutting the interest rate, contrary to Volcker's announcement. The second one was the deepest recession since the Great Depression, because Volcker stuck to his guns and maintained high interest rates to cut inflation, despite prolonged recession and high unemployment.



Initial Recession (grey) forced Volcker to delay plans to raise the real rate. However, towards the end, coming out of the recession, Volcker raised interest rates sharply, as per his announced intention. This action led to decline in inflation (and ALSO a longer recession). The next graph shows more clearly how the actions of Volcker led to elevated Federal Funds Rate (the basic short term interest rate which governs the economy). The “Real” rate is the interest rate minus inflation; according to theory, the economy is affected by the real rate, not the nominal rate. As announced, Volcker raised and held the real interest rate at around 5% until inflation declined to below 5%, and stabilized at this level.



Note that, unlike Hume's constant conjunctions, Romer deduces causality from just one event. He explains and defends this idea. We will provide further support for this methodology in later lectures.

4: Lessons from Experience

In terms of drawing causal arrows in path diagrams, the lesson from this experience is the following: If taking action A reliably achieves goal G, then $A \Rightarrow G$. Here "Reliably" is subjectively defined. We are never SURE about causality. This is a PRAGMATIC knowledge – we believe a causal relationship holds because it works for us. There are unspecified and unknown background factors which create the causal mechanism – it is possible, even likely, that if some of these background factors were different, our causal mechanism would not operate. Applying the same action in different circumstances may or may not lead to the same result. Judgements about whether environment is sufficiently similar is subjective.

Philosophers have been searching for centuries, in vain, for PROOFS of causality. In this connection, the philosophy of Empiricism championed and popularized by David Hume has been A HUGE obstacle to causality. Learning about causal mechanisms REQUIRES knowledge of the goals of the policy makers. These goals are UNOBSERVABLE – in the MINDS of the planners. As we will show in the next lesson, for different goals, the SAME data provides evidence for different causal mechanisms. In fact, we can learn about goals of monetary policy by reading the notes of the policy makers. This is ESSENTIAL to understanding causality.

Concluding Remarks

Q: Given a collection of variables V,W,X,Y,Z – how can we determine if X causes Y?

A: If a free-agent wishes to achieve a specified outcome $Y=Y^*$, and sets $X=X^*$ and achieves the desired result (or approximately so), then we conclude that $X \Rightarrow Y$.

This is a DIRECT causal relationship if V,W,Z do not change (much) during the time interval between setting $X=X^*$ and observing $Y=Y^*$.

If the time interval is short, other variables don't have time to change, and so we can deduce direct causal relationships

Possibly Indirect Causal Relationships

If the observed values of V,W,Z change significantly over the period of time between setting $X=X^*$ and observing $Y=Y^*$ then these variables MAY BE intermediaries – that is, it is possible that

$X \Rightarrow Z \Rightarrow Y$; Change in X led to change in Z led to change in Y.

Unobserved variables, or variables not considered by modeler, were in a favorable configuration to allow causal effect to occur. For intermediacy, we can test by fixing the value of Z and then trying to achieve same causal effect. If causal effect fails to operate, we might look for background variables which prevented this causal effect from operating.

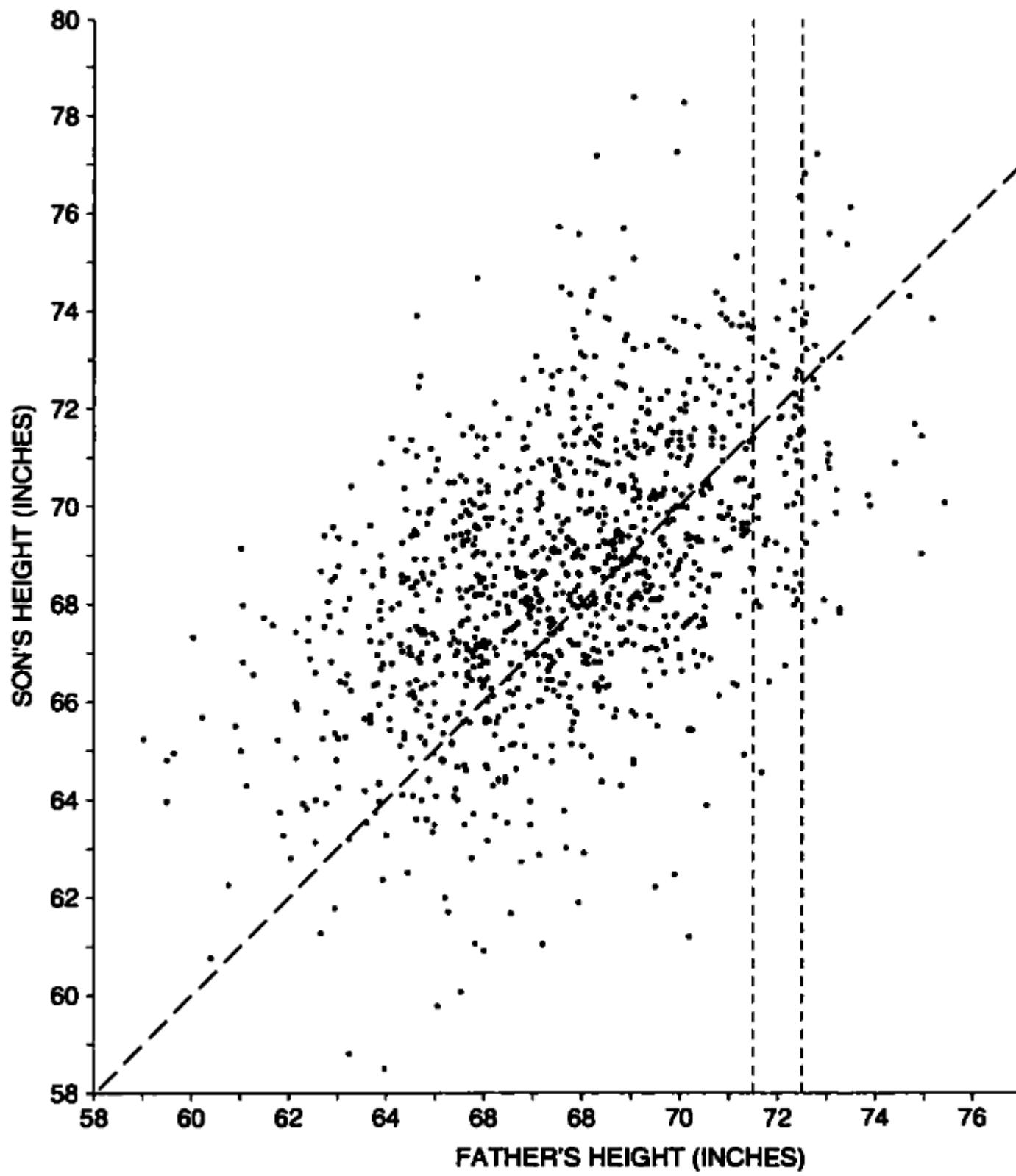
Related Materials

1. “Causality as Child’s Play”: <http://bit.ly/RSRA11B>
2. Writeup of this lecture (Path Diagrams): <http://bit.ly/RSRA11C>
3. Next: “Correlations, Causality & Structure”: <http://bit.ly/RSRA11C>
4. The Incredible Volcker Deflation: Goodfriend & King
<https://www.bu.edu/econ/files/2011/01/GKcr2005.pdf>
5. Wrong-Headed Article about Honig has interesting side-info about Volcker: The Fed’s Doomsday Prophet Has a Dire Warning About Where We’re Headed
<https://www.politico.com/news/magazine/2021/12/28/inflation-interest-rates-thomas-hoenig-federal-reserve-526177>

11D: Common Cause and Correlation

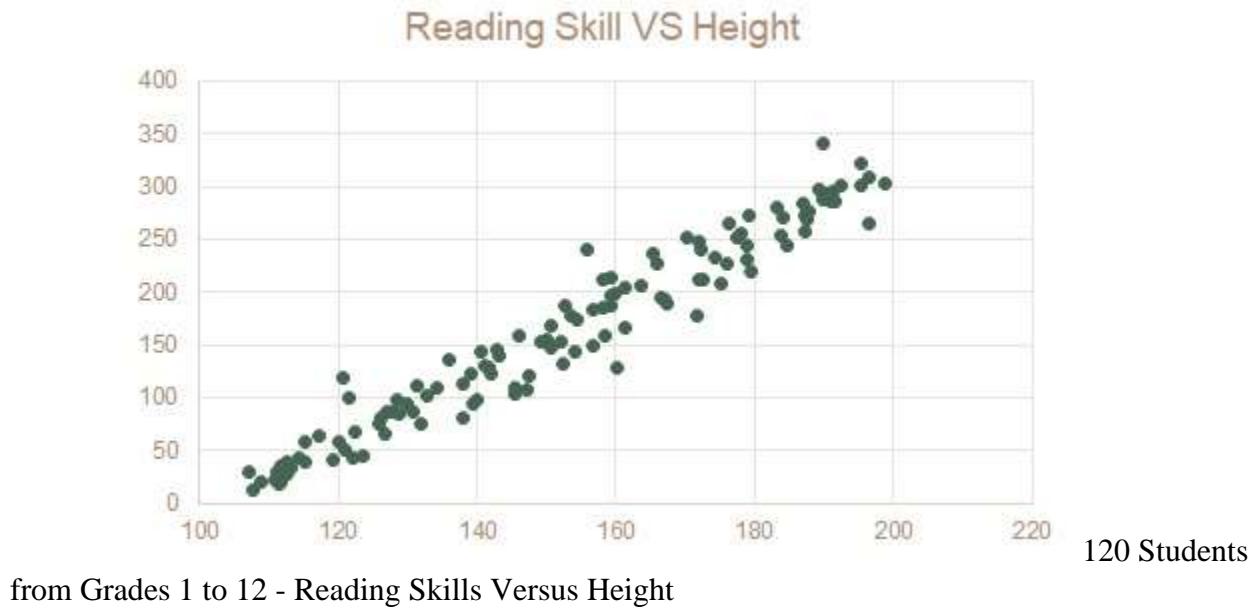
The central problem of statistics is to learn about causation. In this lecture, we will learn some basic methods to distinguish Causation from Correlation. For reasons discussed in an earlier lecture (see also, Chapter 2 of Book of Why by Pearl and Mackenzie), conventional statistics textbooks do not discuss causality in any depth or detail.

One can view Modern Statistics as A search for causality, which went astray! The original goal of Racists (Eugenacists) founding fathers of statistics (Galton/Pearson/Fisher) was to prove the superiority of white races through genetics. As an illustration of genetics, they studied the influence of Father’s height on Son’s height. This is why the following diagram, based on data on 1000+ pairs collected by Pearson, is the starting point of discussions of correlation:



Several setbacks to this original project on statistics of heredity occurred. They discovered the “Regression Effect”: children of taller parents are somewhat shorter, and children of shorter parents are somewhat taller. The distribution of heights of the population remains stable over time – it does not split up into “Tall” and “Short” as it would if taller people kept marrying each other. It was widely realized that causality cannot be established by observational data. Establishing causality depends on what WOULD happen if we changed the values of the cause. But this can never be observed. The data only display correlations, which are symmetric, while causes go in one direction. Because of these considerations, Sir Ronald Fisher came to the conclusion that we must base statistics on CORRELATIONS only. The job of statistics was to summarize the data. The pursuit of science – based on observables alone - required us to ABANDON causality.

After these preliminary remarks, we turn to the topic of this lecture. The following graph, based on artificial but realistic data, shows the strong influence of heights on reading skills in a population of school children from grades 1 to 12:



The graph shows a strong relationship – as heights increase, reading skills also increase. So we can ask: “Does Height Affect Reading Abilities?”. The answer is no – we have designed the data set so that there is no relationship between heights and reading skills. This lecture explains why there appears to be such a strong relationship in the data, and how we can avoid being deceived by this correlation, which is not causal.

This artificial data set on 120 Students, with 10 each from grades 1 to 12, was generated as follows. We created data on heights via the EXCEL formula:

$$\text{Age} = \text{RANDBETWEEN}(12*(4+\text{Grade})-3, 12*(5+\text{Grade})+3)$$

$12*(4+\text{Grade})=60$ for grade 1, 72 for grade 2, and so on. The age of the ten students in grade 1 is chosen to be $\text{RANDBETWEEN}(60-3, 72+3)$. This function generates a random number

between 57 and 75 as the age in months of the student. We populate each of the 12 grades with 10 students, and assign them ages in months randomly so that Grade G has students aged $4+G$ years, allowing for 3 months extra on both sides. The age in months is chosen randomly between $57=60-3$ and $75=72+3$ so that all numbers within the range are equally likely.

Next, for each student, we generate the Height from:

$$\text{Height} = 100 + 7.5 * \text{Grade} + 0.6 * (\text{Age} - [12 * (4 + \text{Grade})]) + \text{Random Variation}$$

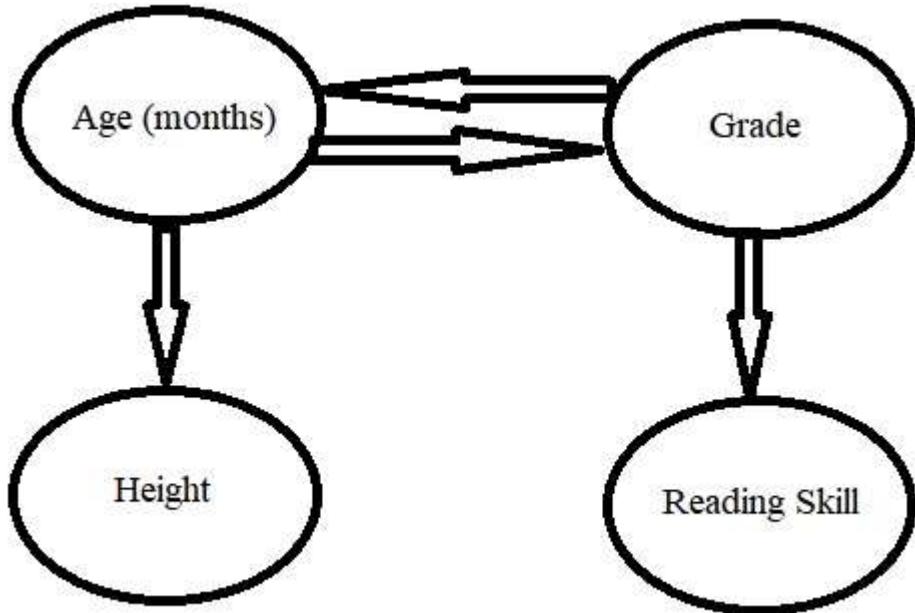
This gives every student a base height of 100cm and adds 7.5cm for each grade. In addition, starting from the base entry age (60 months for grade 1), each additional month adds 0.6 cm to the height of the student. Finally, to account for natural variation in heights, a random number is generated from a standard normal distribution and added to this number to create the height data for the student.

The Reading Skill of each student is generated from:

$$\text{Reading Skill} = 25 * \text{Grade} + \text{LN}(1 + \text{Grade}) * 10 * \{\text{Normal Random Variable}\}$$

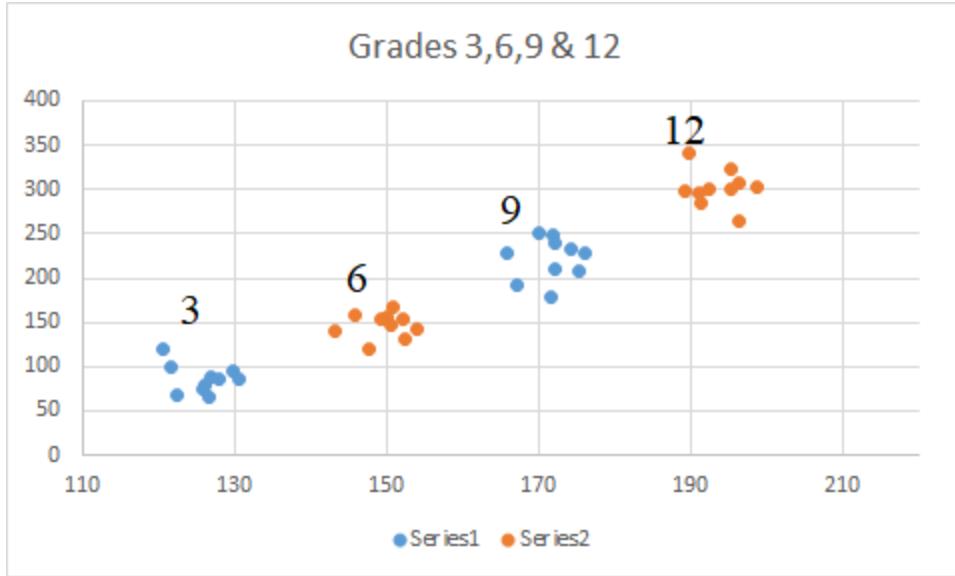
Each grade adds 25 points to the score for Reading Skill. In addition, a random variable is added to this score, to capture normal variation in reading skills. This random variable is scaled by $\text{LN}(1 + \text{Grade})$ because variation in reading skills in the higher grades does not go up proportionally with the grade. This point is not important for our present topic; students may ignore it.

This method of artificially creating data is called a **SIMULATION**. One of the main advantages of simulation is the we KNOW the causal relationships for the data – this is often not the case with real data. The method for generating the data is based on the following causal relationships:



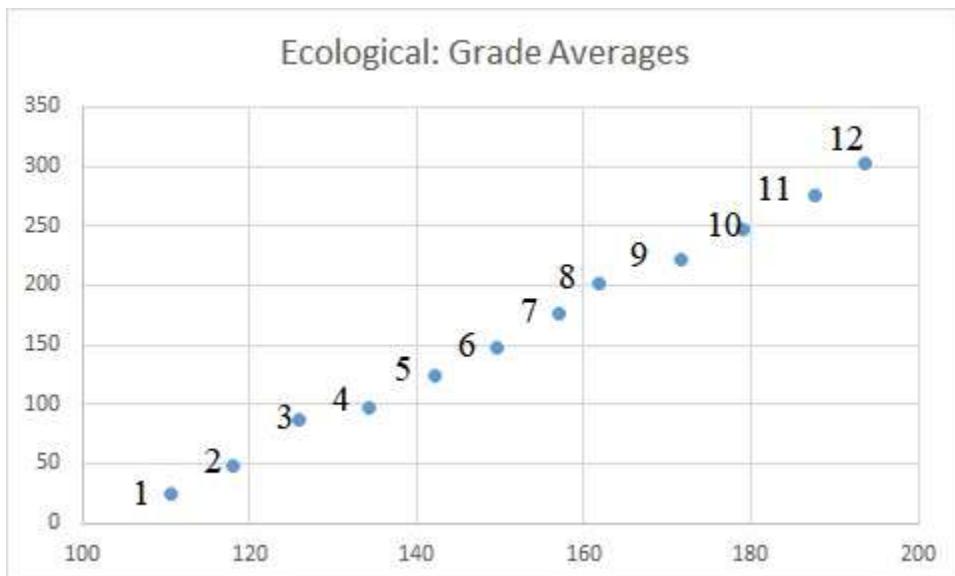
The simulation has been set up so that age is determined on the basis of the grade. In the real world, we might consider age as determining the grade. For the arguments to follow, the causal direction of this arrow does not matter. The above causal diagram shows that the GRADE is a confounding variable – it affects both the Height (H) and the Reading Score (RS). This creates a correlation between H and RS even though there is no causal relationship between them.

The initial graph of H and RS shows a strong relationship because it does not take the confounding variable of GRADE into account. In order to learn that there is no causal relationship, we must “Control” for Grade/Age. This means that we should compare height & reading skill for a COMMON age group, or within a single grade. In this case, there should be little or no correlation, if age is a confounder. In the graph,



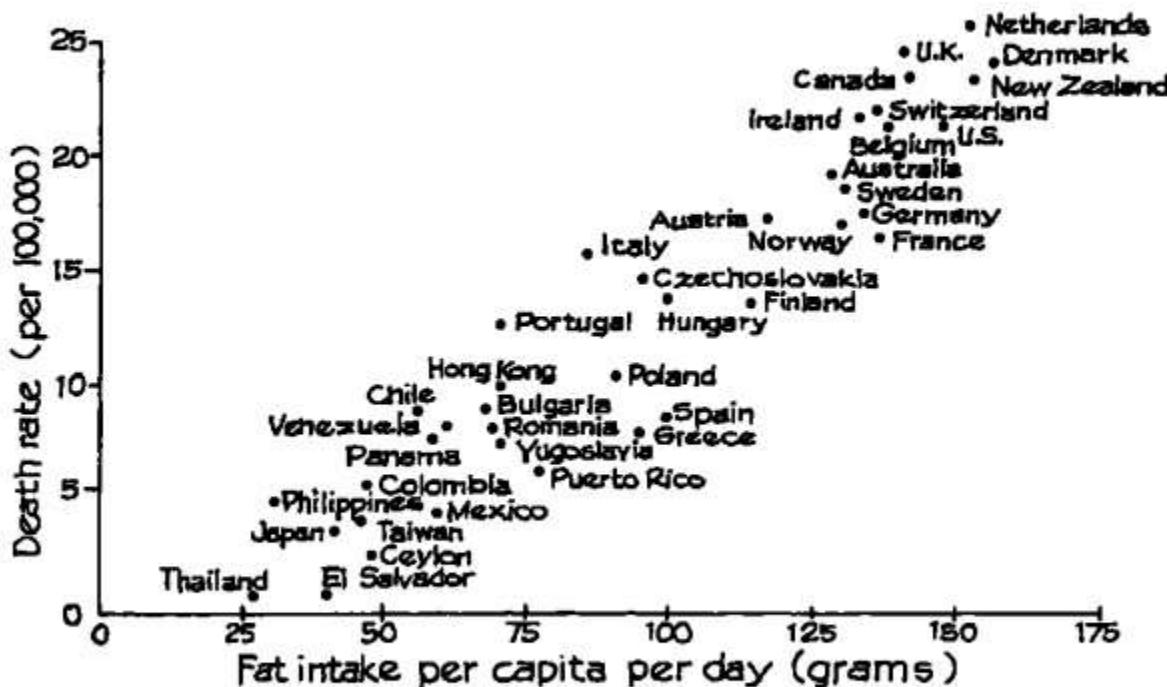
We plot the Heights and Reading Skills for each grade separately – the plot shows the four data sets for grades 3, 6, 9, and 12. Within each class, it is clear that the cluster of points shows no relationship between Height and Reading Skill. Controlling for Grade solves the confounding problem, and reveals the lack of causal relationship between H and RS. It is worth noting that the above graph is based on controlling for GRADE. We would get similar results by controlling for AGE, but that would require separate calculations, since students of similar age can be in different grades (by design of simulation).

The problem that H and RS appear to be correlated across grades is also known as “Ecological Correlation”. This is correlation based on GROUP averages, rather than INDIVIDUALS. It is well known that this can be misleading. Even if there is a genuine correlation, based on causation, it appears to be much stronger at the group level than it is for individuals. This is because individual level variations disappear at the group level.



As another example of this phenomena, consider the following graph of Fat-in-Diet & Breast Cancer, considered at the group level of countries. The graph appears to show a strong relationship. Can we conclude that higher levels of fat in diet cause breast cancer? It is worth noting that the evidence for smoking and lung cancer first emerged in this way – increase in aggregate levels of smoking and also in lung cancer. Nonetheless, there are many reasons to doubt that there is any relationship. If we look for confounding factors, we find that higher fat in diet is associated with higher levels of wealth, or GNP per capita. Thus richer countries tend to have higher percentage of fat in diets and also have higher proportion of breast cancers. This leads to a large number of possibilities for confounding factors. Dietary patterns vary with the level of wealth. Generally speaking, sugar consumption is also an increasing function of wealth. So a graph of sugar consumption and Breast cancer would look very similar to this graph. Yet, no one has suggested that sugar leads to breast cancer.

Figure 8. Death rates from breast cancer plotted against fat in the diet a sample of countries.



Note: Age standardized.

Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

What can we conclude about the source of the association (correlation) between fat-in-diet and breast cancer which is displayed in the graph? Increasing levels of wealth lead to

changing patterns of diet. Some of these differences protect against certain types of cancers, but also increase the risk of some other types of cancer. At this point, not enough is known about the risk factors for cancer to come to any clear conclusion about this matter.

We come back to the main topic of this lecture. Given that we observe a correlation between X and Y, how can we tell if it is due to a causal relationship between X and Y, or whether it is caused by a confounder, a common cause of both variables? In the conventional approach to statistics, there is only one way to handle confounders. As pioneered by Sir Ronald Fisherian, Randomized Trials equalize the proportion of confounders across the Treatment and Control groups. However, as we have seen, randomization is not possible in many cases. In general, statisticians remain skeptical about the possibility of controlling for confounders in “observational studies”, when randomization is not possible. We will discuss below three different methods for controlling for confounders, which allow us to distinguish between correlations created by confounders and genuine relationships. All three depend on correctly identifying the confounding variable(s). This requires understanding of the real world context of the data, and can never be done mechanically on the basis of the observed data.

We have already seen one method of CONTROLLING for Confounders. This is to hold them CONSTANT while making comparisons across Treatment and Control groups. In our example of H versus RS, holding the GRADE constant involves analyzing the data separately for each grade. Within any one grade, the confounding variable does not vary by much. So if Height and Reading Skill are correlated, it cannot be due to the common cause of Grade, since Grade does not vary. When we looked at H vs RS within a grade, we found no relationship between the two variables. Similarly, for the Fat-Diet-Example, we could group countries by approximately equal levels of wealth, and plot each group separately, to see if wealth is a confounding variable.

In conventional statistics, the standard method involves calculating the Partial Correlation of Y, X given confounders Z₁, Z₂, ..., Z_k. We will not define exactly what this means, as it involves making a lot of assumptions about the data which do not hold for most data sets. However, if all of these assumptions (of the regression model) are valid, then this partial correlation coefficient can be computed by running regression of Y on X together with the confounders Z₁, ..., Z_k. In this situation, the estimated regression coefficient of X will provide with an estimate of the partial correlation. This method is not very satisfactory because the assumptions can FAIL to hold and lead to wrong and misleading results for many reasons. See Choosing the Right Regressors and A Realist Approach to Econometrics for more detailed explanation.

However, in the present simulation example, the data has been generated according to the assumptions imposed on regression models. Therefore, running a regression of RS (Reading Skill) on Grade and Heights provides us with the right results. The regression shows that Grade is highly significant as an explanatory variable, while the Height does not matter – it is not significant. The regression results are as follows:

$$RS = 7.5 + 25.4 \text{ Gr} - 0.06 \text{ Ht},$$

where 25.4 has an extremely small p-value, while the coefficient of Ht is negative. This shows that the partial correlation of Gr with RS is strong and significant. The partial correlation of Ht is actually negative, but the p-value is high, meaning that it is not significantly different from zero.

The first method, of holding the confounder constant, works well if sample size large enough, and we get decent sample sizes within each subgroup where the confounder is constant. However, there are many cases where we end up with unbalanced small samples within different subgroups. In such situations, it is hard to see whether or not there is correlation with subgroups and to make a decision about the overall picture, whether or not the correlation between X and Y is caused solely by the confounder. In such situations, there is an ALTERNATIVE Strategy: Adjust variables for confounder. We will now explain how this method works for the artificial data set on Heights and Reading Skills.

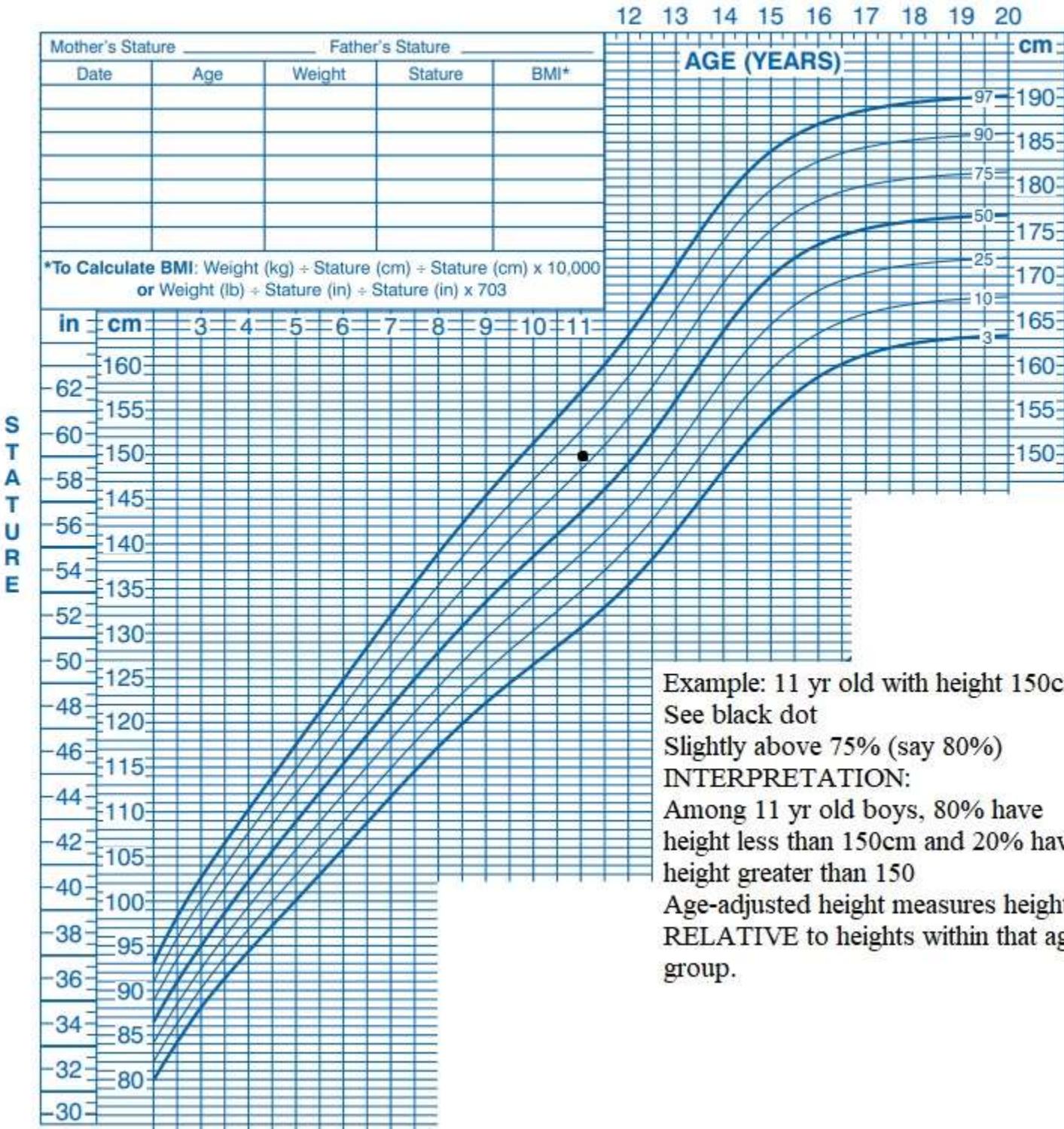
The idea of a Grade-Adjusted Reading Score is very natural. We do not have (and cannot have) a common reading skill test for all 12 grades. Instead, we give each grade its own, level-appropriate test. These test scores are not comparable across grades. We can SET the test scores so that 100 reflects median (“average” or “normal”) performance within each grade. A score above 100 signifies a better than average performance, while below 100 is below average. This Grade-adjusted score removes the effect of the confounding variable Grade from the Reading Skill score.

Similarly, the idea of an age-adjusted Height is also very natural. Pediatricians use charts like the one below to assess whether the growth of children is within normal ranges. We illustrate the use of this chart to see how it gives us an age-adjusted height. Suppose an 11 year old has height of 150cm. This data point appears as a black dot on the chart below. The middle solid blue line is the median, or the 50%. Half of the population is above, and half is below this line. The median line is around 142cm, while the 75% line is around 148cm. As a rough visual approximation, it seems that 150cm is around 80th percentile among 11 year olds. Thus the age-adjusted height score can be taken to be 80.

2 to 20 years: Boys Stature-for-age and Weight-for-age percentiles

NAME

RECORD # _____



Example: 11 yr old with height 150cm
See black dot

Slightly above 75% (say 80%)

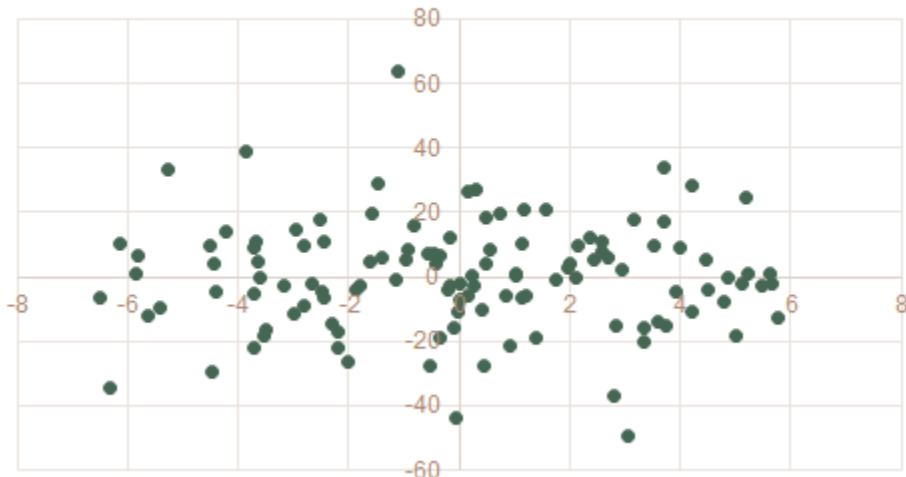
INTERPRETATION:

Among 11 yr old boys, 80% have height less than 150cm and 20% have height greater than 150

Age-adjusted height measures height RELATIVE to heights within that age group.

This means that 80% of 11 year olds are below this height, while 20% are above this height. The age-adjusted height score in percentiles is comparable across ages. The effect of age has been removed from this score. Now if we plot the age-adjusted variables, any relationship between H and RS cannot be due to the confounding variable of age. For our artificial data set, the GRADE adjusted Height and Reading Skill shows no relationship (see graph below). The Grade-Adjusted Heights range from -6 to +6. These are cm of difference from the average height within the grade of the student. Similarly, the Grade-Adjusted Reading Skill ranges from -60 to +60. Again, these scores are the difference from the average score within a single grade. The graph shows NO CORRELATION. We can see that by noting that the range of the Reading Skill scores remains the same across all Height ranges. The height has no affect on reading skill after we adjust for Grades.

Grade Adjusted: Height vs Reading Score



Some remarks are in order. The adjustment for grades was done by first computing the average height and average reading score in each grade. Then, for each student, we subtracted the average height of the grade from the students height to adjust the height for grade. Similarly, the average score for Reading Skill within a grade is subtracted from an individual score to give the grade-adjusted Reading Skill. We could also do adjustment for age – these would be somewhat different, because we would group students together by age (in months) and then do the adjustment. The end-results would be similar to the graph above, and show that there is no correlation between height and reading skill after adjusting for age.

A second important point is that we get clean results because the simulated data sets satisfy clear cut causal assumptions. Real Data Sets are often messier. In REAL data sets, height CAN be causal influence on reading skills. This can happen if low height is caused by malnutrition and stunting which would also result in low cognitive skills. It can also happen through increased self-esteem and confidence which has been found to be significantly associated with height. Discovery of causes and of confounders requires real world knowledge, thought, data gathering, and analysis or experiment. It is by no means a mechanical process.

Summary & Conclusions

We will recapitulate the main lessons from this lecture, which covers many points which are not part of mainstream statistics. The goal of real statistics is the discovery of causal connections between variables. Causal connections are hidden – they are never observable. Causal relations imply patterns of correlation about observable data. We can get support (but never proof) about a causal hypothesis by checking the implied pattern of correlations. We can also disprove causal hypothesis if we find correlations which do not correspond. Causal hypotheses come from knowledge about the real world, supported by observed correlations.

If we observe a strong correlation between two variables X and Y, we must look for a causal explanation for this correlation. The simplest explanations are based on direct causation, which can be $X \Rightarrow Y$, or $Y \Rightarrow X$, and sometime mutual or bi-directional causation as in $X \leftrightarrow Y$. However, there are many cases where correlation can exist without any of these three causal relationships. An important case where this happens is when both X and Y have a common causal factor Z. This can be pictured as $X \leftarrow Z \rightarrow Y$. This is called a “fork” causal pattern. Variations in Z are transmitted to both X and Y and hence lead to a correlation, even though there is no causal link between X and Y. In this situation, the common cause Z is also called a confounder.

How can we discover if an observed correlation between X and Y is due to a confounder? The first step is to learn about potential confounders; this can only be done on the basis of real world knowledge about X and Y – it can never be done by any kind of data analysis confined to the observations on the variables X and Y. If we have a variable Z about which we think that it may be a common cause, we can assess this by conditioning on Z. This means that we hold Z constant and then study the relationship between X and Y. This is called controlling for Z and can be done in several different ways. If the correlation between X and Y is solely due to the common cause Z, then this correlation will disappear if we hold Z constant. Thus examining association between X and Y when Z varies very little will tell us whether or not Z is a common cause. Another way to assess this is to adjust variables X and Y for the common cause Z. This involves removing the influence of Z from X and Y. Z-adjusted values of X and Y should be independent, if the correlation is solely due to common cause Z. The third method is based on conventional statistics and involves regressing Y on X and Z. If the coefficient of X is not significant in this regression, then Z may be a common cause of X and Y. This third method works under some very restrictive assumptions required to run a valid regression. That is why we do not discuss this method in any detail in this lecture.

11E Causation as Deep Structure

One of the central messages of “Real” Statistics is that data cannot be analyzed without knowledge of the real-world structures which generate the data. In particular, correlation is a surface phenomenon, while causation is a real-world phenomenon, so the two are radically different. Furthermore, observational data can NEVER give us certain conclusions about causality, especially the experience-based and goal-driven causality concept that we have been studying. In the previous lecture, we saw that a common cause ($Z \Rightarrow X$ and $Z \Rightarrow Y$) can create

strong correlation between two variables X and Y, without any causal relationship between the two. If Z is unknown or unobserved then correct causal inference is impossible from observations of X and Y. However, if we can MANIPULATE X then we can reject the causal link $X \Rightarrow Y$ suggested by the data, because we will see that setting values of X has no effect on Y. The goal of this lecture is to demonstrate that Causal Effects CANNOT be inferred from OBSERVATIONS. To do this, we will create an example where the underlying real-world causal mechanism which generates the data remains the same throughout. However, some peripheral changes in data gathering mechanism cause massive difference in the “observed” causal connections. That is, vastly different correlation structures can represent the same underlying causal mechanism. From this we deduce that it is necessary to do a deep investigation into the real-world mechanisms, in order to learn about causality – we cannot do this by looking at the numbers alone, without knowledge of their real-world context

A Simple Economic Model

Since some knowledge of real world is necessary to understand causal mechanisms, we introduce some basic economic concepts. Assume for simplicity that there is only one interest rate and one inflation rate throughout the economy. If Inflation (Inf) = 10%, PKR 100 today is equivalent to PKR 110 next year. A Nominal interest rate (NIR) of 10% will give you PKR 110 next year for a payment of PKR 100 this year. SAME amount of money (in terms of real value = purchasing power). So REAL interest rate is 0%. This leads to the definition of the Real Interest Rate (RIR) as follows: $RIR = NIR - Inf$. Economists agree that it is the real interest that people consider when making decisions about money, and so this is what affects the real economy. Assuming that this is so, it would make sense for the Central Bank to set a target for the Real Interest Rate and to try to achieve this target by setting the nominal interest rate. One theoretically desirable target for RIR is to have it match the real growth rate of the economy. To simplify the argument which follows, we will assume that the real growth rate is fixed at a constant value of 3% and this is known to everyone. Furthermore, we will assume that the Central Bank (CB) aim to use monetary policy to arrive at real interest equal to 3% at all times. This then is the underlying causal mechanism which drives interest rates.

Next we consider how inflation works in our hypothetical economy. We assume that this is a small open economy which produces only one product – say bananas, or oil. This is sold on the global market, and the revenue is used to import all other goods. The prices of exports and imports are determined on the global market, and have no connection with anything happening within the small domestic economy — this makes the prices “exogenous”. This also means that inflation is exogenous. We assume that inflation changes from month to month on the basis of random shocks to global prices:

$$Inf(t) = Inf(t-1) + RSh(t)$$

The Central Bank meets on a monthly basis, and attempts to measure current inflation, in order to set the nominal interest rate equal to $Inf(t)+3\%$, so as to arrive at the policy goal of 3% Real Interest Rate. However, current inflation rate is hard to measure because all of the relevant

data required to measure it is not readily available at the current time. As a result, the CB measures current inflation with a Measurement Error. Let us say that $\text{Inf}^*(t)$ is the CB measure of $\text{Inf}(t)$ available at the time the decision to set the interest rate is being made:

$$\text{Inf}^*(t) = \text{Inf}(t) + \text{MER}(t) \quad \{ = \text{Inf}(t-1) + \text{RSh}(t) + \text{MER}(t) \}$$

We have taken causality to be actions driven by goals. In terms of this definition, the fundamental causal mechanism is the policy decision driven by the goal of achieving 3% real interest. We will show that different assumptions regarding the random shocks (RSh) and the measurement errors (MER) will lead to very different patterns of surface correlations, even though the fundamental causal mechanism driving the system remains the same throughout.

What the Data Shows:

Case 1: Suppose that the CB has no information regarding the current value of $\text{Inf}(t)$, but it can measure the last month's inflation perfectly. This can be represented as $\text{MER}(t) = -\text{RSh}(t)$ — the measurement error is exactly the negative of the random shock so that the CB's best guess at current value of Inflation is simply last month's inflation. In this case, CB will set the nominal interest rate as follows:

$$\text{NIR}(t) = \text{Inf}^*(t) + 3\% = \text{Inf}(t-1) + 3\%$$

This will cause the Real Interest rate to deviate from the desired target of 3% by the amount of the random shock:

$$\text{RIR} = \text{NIR}(t) - \text{Inf}(t) = 3\% + \text{Inf}(t-1) - \text{Inf}(t) = 3\% - \text{RSh}(t)$$

Anyone who studies data generated according to these mechanisms will come to the following conclusions – based on correlations:

1. The rate of inflation at time $t-1$ determines the nominal interest rate at time t .
2. Inflation rate and nominal interest rate are strongly positively correlated.
3. The Real Interest rate has no correlation with any domestic economic variable (since it is determined by exogenous global shocks to global prices).

As statements about the data correlations, all three are TRUE statements. As statements about the underlying causal mechanisms at work, all three are FALSE statements. WHY are these three assertions false? To understand this clearly, we have to distinguish between surface appearances and underlying caused in the real world.

How Observed Correlations Differ from Underlying Causation

FIRST: it is not true that the rate of inflation in the past period determines the current nominal interest rate – this is only appears to be the case because the central bank cannot measure the current inflation rate and hence approximates it by the past inflation. Suppose that the measurement error disappears; the Central Bank finds new ways to get accurate estimates of the current inflation. On the basis of correlations, in the initial scenario, we will say that $\text{Inf}(t-1)$ determines $\text{NIN}(t)$. Given $\text{Inf}(t-1)$, $\text{Inf}(t)$ will not add anything to the explanation of $\text{NIN}(t)$ so it

would be regarded as irrelevant to the determination of current NIN. However, after increase in accuracy of measurement, the correlations will say that $\text{Inf}(t-1)$ is irrelevant to determination of $\text{NIN}(t)$, and only $\text{Inf}(t)$ matters for this purpose.

SECOND: The strong positive correlation of Inf and NIN is an accidental byproduct of the real causal mechanisms driving the system. To see this, suppose that Inflation adversely affects growth. To take an extreme (implausible) case, suppose that the growth rate is exactly equal to negative of the inflation rate. Since the Central Bank desires to set the real interest rate equal to the growth rate, all it has to do is to set Nominal Interest Rate to ZERO. In this case:

$$\text{RIR}(t) = \text{NIN}(t) - \text{Inf}(t) = 0 - \text{Inf}(t) = \text{Growth Rate of Economy}.$$

The policy objective of setting real interest rate equal to the Growth Rate would be achieved by setting Nominal Interest equal to 0. Suppose that for a period of time, growth rate is independent of inflation. In this scenario, we will see a strong positive relationship between nominal interest rates and inflation. Later, suppose that due to structural changes in the global economy, inflation starts to adversely affect growth. Then we will see that the correlation between inflation and Nominal Interest declines substantially (to 0% in the extreme scenario that growth is negative inflation). However, the underlying causal mechanism driving monetary policy remains the same throughout.

In the original setup, the correlations suggest that $\text{Inf}(t-1)$ causes $\text{NIN}(t)$. If we eliminate measurement error, the correlations say that $\text{Inf}(t)$ causes $\text{NIN}(t)$. There is a third possibility. Suppose that the CB can accurately forecast $\text{Inf}(t+1)$ and wants to set $\text{NIN}(t)$ to offset future inflation in the next month. This is sensible, since the rate it sets will actually be operative in the next month. THEN the correlation structure will show a strong positive correlation between $\text{NIN}(t)$ and $\text{Inf}(t+1)$, leading observers to believe that setting the policy rate determines inflation, and high rates lead to high inflation, even though the reality is far different.

THIRD: In the system as described, the data cannot provide us with any information about the impact of the real interest rate on economic variables. That is because the Central Bank is keeping the real interest rate fixed. We need variations in the interest rate to see what the affects of changes in interest rate are. But, one of the central insights of Judea Pearl is that SETTING the real interest rate can have very different effects from OBSERVING the interest rate. This is worth understanding in greater depth and detail.

Suppose that, in the system as originally described, the random shocks are quite large. Then the Real Interest Rate, which is $3\% - \text{RSh}(t)$, will fluctuate quite a bit. We could study the impacts of RIR by isolating those periods where the RSh is large and positive, and comparing them with those periods where the RSh is large and negative. A comparison of these two periods would give us the OBSERVED impact of changes in real interest. However, suppose that agents in the economy IGNORE these fluctuations, knowing that these are random changes which will average to zero in the long run. The differences in periods where RIR was high, when compared with periods where RIR was low, will reveal nothing meaningful. HOWEVER, now suppose that the Central Bank announces that it plans to RAISE the Real Interest Rates and HOLD them at this high value for an extended period of time. This time the agents will pay attention to the high

interest, and knowing that this change will persist, they will change their plans. Then we will be able to observe the effects of a goal-directed SETTING of the real interest rate to a high value, and these will be quite different from OBSERVING the effects of high interest rates in past. This differs from both Pearl and Woodward in making the goals of the agent central to the causal mechanism.

Causality is in the Model, Not in the Data

Our Action-Goal approach to causality says that we need to take into consideration the goals of the agents. In order to link actions to goals, the agent must have a model, however primitive or sophisticated, about causal consequences of actions. As we have seen, data correlations provide no guidance about causality, but causal models do. Since agents wish to achieve goals and seek to do so, they must use some causal model to predict the effect of their actions. In addition to the agents model, there is a TRUE structural model which described the real-world causal mechanisms which link actions of agents to their consequences. This model may be infinitely complicated and never fully known or understood. Systematic deviations of agents model from reality will be noticed in discrepancies between anticipated effects and observed effects of actions. Such deviations will lead to attempts to improve the model so as to produce a better match.

Understanding the causal mechanisms which drive the world we live in requires thinking about both the agents model and the true model governing real-world causality. In case of monetary policy, reading through the minutes of the monetary policy meetings gives a good amount of insight into both the goals, and the models, of the policy makers. Understanding the real world causal mechanism requires different kinds of efforts and investigations.

Concluding Remarks:

In this lecture, we showed that observed correlation structures can provide extremely misleading clues about the underlying causal mechanisms. An empiricist/positivist mindset which has been dominant in Western intellectual tradition since David Hume has blocked the path to deep investigations of unobservables, on the grounds that this is not scientific. Human goals and human experiences are deeply unobservable – my life experiences cannot be observed by others and cannot even be communicated to others. However, these experiences are the foundations on which our knowledge is built. By excluding personal experiences from scientific knowledge, we lose the possibility of understanding causality.

In the example that was studied, the observed correlations suggest that inflation and nominal interest are positively correlated. With slight variations, the data can show that $\text{Inf}(t-1) \Rightarrow \text{NIR}(t)$ or that $\text{NIR}(t) \Rightarrow \text{Inf}(t+1)$, but neither causal implication is true. The OBSERVED correlation between NIR and Inf does not provide us with any clue as to what would happen if we SET the NIR to achieve different goals from the ones currently in use. This remark goes one step beyond the insight of Judea Pearl, who has cogently argued that SETTING a variable has different impacts from OBSERVING a variable. It is not just setting the variable which matters, but the GOAL for which the variable is being set makes a huge difference. This goal is always

unobservable in the statistics, though we can learn about the goals of monetary policy by studying the discussions which take place in making the policy decisions.

This example demonstrates how it is necessary to go beneath the surface and explore the unobservable structures of reality which generate. Without doing so, it is impossible to learn about causal mechanisms.

12: Simpson's Paradox

Blurb for the 12th and final chapter

12A: An Admissions Paradox

The goal of this lecture is to present, by example, a new theory of causality: goals drive actions that lead to observed consequences, which may or may not match the goal (intended consequences). The actions and observed consequences are observable, but the intended consequences are not. Human experience consists of knowledge gained by the mismatch between the observed and intended consequences, and modifications of actions to create a better match. All of this is lost in the standard approaches to causality, based on the observables only.

Open-And-Shut Case: Discrimination

We consider admissions data for a mythical “Berkeley University”:

1000 Female Applicants 900 Admits

1000 Male Applicants 100 Admits

With a 90% admit ratio for females, and a 10% admit ratio for males, it seems crystal clear that Berkeley discriminates against males. The numbers can't lie? Or, can they?

Some Theory and Philosophy

When we say that Berkeley discriminates, we are making a claim is being made about a GOAL of a decision-maker; an action taken to favor females. As should become clear after this lecture, GOALS CANNOT be deduced from data. When you try to do so, paradoxical results OFTEN emerge. The data can easily point to directions different from the unobserved goals. The theory of causality that we propose to examine in this lecture sets up the following causal sequence as basic to the study of causality:

Goals => Action => Consequences

The current best account of causality is given by Woodward in “Making Things Happen.” He argues about causality in terms of how different actions have different consequences, and if an action reliably results in an effect, this is causality. But Woodward’s account pushes into the background the Agent and his goals: WHO makes things happen, and WHY? Philosophers deliberately AVOID thinking about these CENTRAL questions, because they reject

“Anthropocentric” explanation: causality exists whether or not there are human beings. So, accounts that make human goals central cannot account for more general types of causal effects, which would exist even if no human beings were around. Also, they are wedded to positivism – the idea that only observables matter for science. Human intentions & goals are NOT observable, and hence not suitable to incorporate in scientific theories.

Real Statistics is founded on a new theory of knowledge (Epistemology). We argue that ALL knowledge is anthropocentric – We cannot aspire to Godly status. Human Experience is the basis of all knowledge. Superficially, this also appears to be the claim of “empiricism”, the dominant Western philosophy of knowledge, starting from David Hume. But Hume takes knowledge to be concerned only with the external world, or objective reality. In contrast, we take human knowledge to be built upon as our internal psychological experience. This makes our knowledge SUBJECTIVE, LOCAL, EPHEMERAL, not UNIVERSAL. According to standard Western theories of knowledge, subjective human experience does not classify as knowledge at all.

A central aspect of our experience is that of AGENCY: Our goals influence our actions, which lead to outcomes, often different from our intended goals. This leads to REFLECTION on why our actions did not create the desired effects. We create THEORIES about the world which link actions to consequences, and use them to decide on how to modify our actions to better achieve desired results. This is called “learning from experience.” But our theories play a crucial role in this learning. IDENTICAL experiences can lead to DIFFERENT lessons for different agents, depending on the theories they use to analyze the experience. For one agent, a sequence of six failures can be interpreted to mean that “I am incapable of achieving success” and lead to the abandonment of further effort. For another agent, the same six failures might be interpreted to mean that “I need to try harder” and lead to renewed efforts.

Back to the Admissions Paradox

We have displayed data that APPEARS to show an open-and-shut case against Berkeley for discrimination against males. BUT, the data cannot show the GOALS of Berkeley, and the REASONS why the admissions ratio for females is high, and that for males is low. To understand goals and reasons, we have to look at the agents – the decision-makers. There are two sets of decision-makers. Let us first focus on the students. The goals of the students are to get admission, and the action they take to achieve the goal is the application process. But let us look more deeply into these goals. When applicants apply to Berkeley, what are their goals? SUPPOSE that investigation reveals the following:

- Females seek admission into the Literature program (Lit).
- Males seek admission into the Engineering program (Eng).

Suppose now that

- Literature has a 90% admit rate: 1000 Female Applicants => 900 Admits
- Engineering has a 10% admit rate: 1 000 Males => 100 Admits

It should be clear that there is NO DISCRIMINATION on part of Berkeley. The two types of candidates have different goals – Females want a Lit degree while Males want an Eng degree. Lit has an easy admissions policy, while Eng has a difficult admissions policy. The difference in goals accounts for the difference in application strategy (Lit vs Eng), which accounts for the difference in admit ratio. The two populations are not comparable in terms of goals.

Simpson's Paradox

The famous Simpson's Paradox can be illustrated by a small extension of this example. First, mix up the goals of the two genders a little bit. Suppose that 90% of the females apply to Lit and 10% apply to Eng. Conversely, 10% of Males apply for Lit and 90% apply for Eng. Both departments have gender-blind admissions, so 90% of applicants to Lit get in, and 10% of applicants to Eng get in. This yields the following numbers:

1000 Females: 900 => Lit => 810 Admits, 100 => Eng => 10 Admits

1000 Males : 100 => Lit => 90 Admits, 900 => Eng => 90 Admits

Male Admits 90+90 :=: 18%

Female Admits 810+10 :=: 82%

According to the admit ratios (18% for Males, 82% for Females), Berkeley University discriminates against males. BUT: Both Departments have gender-blind admissions! The paradox arises from not looking at GOALS – Actions – Outcomes. If we lump together students who have different goals and take different actions to achieve their different goals, we mix together two separate populations, and arrive at the wrong conclusion.

We can also change the data so that both Departments FAVOR males over females, but the data shows that the University FAVORS females. Suppose that, in view of the few males in the department, Literature encourages males, and admits ALL males 100% (as against 90% admits for females). On the other hand, Engineering is dominated by male chauvinists, and rejects all females sight unseen. Then the numbers would be:

Females: 900 => Lit => 810 admits 100 => Eng => 0 admits

Males: 100 => Lit => 100 admits 900 => Eng => 90 admits

Both departments discriminate HEAVILY against females, but the data shows that the UNIVERSITY discriminates HEAVILY in favor of females! The point is that the University is NOT a decision-making AGENT, so considering what the university does is problematic. The departments are the agents, and have admissions goals, so they must be considered separately.

Resolving the Paradox using Agents and Goals

From the last data set, we learn that Eng admits 10% Males and 0% Females, and Lit admits 100% Males and 90% Females. BOTH departments discriminate against females. We learn this from the NARRATIVE, not from the data. The narrative says that Lit would like to

encourage the male applicants, because it is overwhelmingly female, while Eng discriminates against females because of male chauvinism. Without this narrative, we cannot conclude that there is discrimination in the departmental admissions (either for or against males). The data cannot tell us about the goals. We can easily come up with different narratives which will generate exactly the same data, but will show that there is no discrimination by gender at either department. For a very simple such narrative, suppose that all males have SAT scores of 1200 and all females have SAT scores of 1000. Admission is gender-blind at both departments, based solely on SAT scores. If Eng chooses the top candidates, it can end up choosing all males. Similarly, if Lit chooses to make the cut, it will end up choosing all the male applicants. This re-inforces the main lesson of real statistics: we must go beyond the data to understand the real-world mechanisms which generate the data.

In particular, correct understanding of causal mechanisms requires understanding the goals of the agents, and how they act to achieve these goals. Exactly the SAME data would have different interpretations if actions and goals are different. In understanding causation, we must also take into account the STRUCTURE of the world, or Environmental Variables. These also have very important effects on causality. For example, details of HOW the admissions process is carried out at Berkeley, in the two departments, are very important in the assessment of discrimination. For example, suppose that the admissions process is mechanical, based solely on a weighted average of three numbers – SAT Math Score, SAT English Score, and Grade Point Average. Then, regardless of percentages, there is no gender discrimination (interpreted in terms of goals of the admissions committee). We will now give some more examples to show that exactly the same data, with different underlying real-world structures, can be based on radically different causal mechanisms. This reinforces the idea that we must understand real-world structure to understand causality.

As an example of a different environmental structure, suppose that Berkely Admission Office admits all students (Lit or Eng) via a single, unified process. Students choose majors AFTER being admitted. In the university-wide admissions, we find that 1000 Females applicants lead to 810 female Admits, while 1000 Males Applicants lead to 190 male Admits. Students CHOOSE their majors so that Literature attracts 810 Females and 100 Males, while Engineering attracts 90 Males only. This is exactly the same data as before, but now the interpretation must be radically different. Male chauvinism can no longer be the answer to the question of “Why is Engineering 100% Male?”. To find the correct answer would require deeper research into the reasons why males and females make these choices. ONE simple explanation could be based on the agency of the Male and Female students: Females prefer Literature while Males prefer Engineering. But many other explanations are possible. The Literature and Engineering Departments MAY influence these choices in many ways. ARE the Departments AGENTS? Do they influence the choices of the students? Alternatively, what are the other factors involved in shaping choices made by the students?

The Agency of the Admissions Department

Since data and reality can point in dramatically different directions, it is useful to define Discrimination in terms of Goals, rather than statistics. In particular, we will say that gender

discrimination occurs when the goal of Admissions is to admit more males (or females). Now suppose we observe a highly imbalanced admissions ratio, like the first example: 90% admits for females and only 10% admits for males. When we look into the admissions process, we find that it is gender blind: the admissions committee looks at data that does not provide any clues as to the gender of the student. Then we can eliminate discrimination (goal-based definition) as a cause of this imbalance. Instead, we must look for other causes. One possible explanation might be that admissions are strongly linked to “reading scores”. It turns out that reading scores for females are higher than those for males in all countries. This is just a hypothetical explanation to show that details of the admissions process matter in determining causal effects.

Another important issue to note is that the use of SAT scores does not eliminate agency! The Admissions Department can choose other criteria. Average SAT Math scores of males are substantially higher than those of females in the USA. This is a cultural artifact since the reverse holds in some other countries. In any case, use of the SAT math is biased against females, while SAT English is biased against males. To learn about discrimination, we must ask the deeper question of “How/Why” were these criteria chosen by the admissions department?

Suppose that we ask the Admissions Committee about their goals, and they respond that our Goal is to create successful educational outcomes. Then we must ask “What is the meaning of success?”, how they measure success, and how they think that the admissions criteria chosen are helpful in achieving these. A wide variety of answers are possible, and each would create different types of causal mechanisms linking gender to admissions.

More Agency Issues

When we expand the scope of causality to include the goals of the decision-making agents, many more questions arise about this (hypothetical) example. For example, “Why are there 90 people in Eng and 900 in Lit?”. It could be that Berkeley wants to specialize in Literature, and keep Engineering small. Alternatively, it is possible that Engineering Dept is mediocre, and few students want to go, preferring to go to better quality departments elsewhere. Or, Engineering is excellent, and deliberately small, for exclusivity? ANSWERS to these questions about causes are NOT in the DATA. They lie in the GOALS and INTENTIONS of the Agents, and the STRUCTURE and ENVIRONMENT which creates causal effects of actions of agents.

An aphorism that expresses a half-truth is: “Causality is in the MODEL, not in the DATA”. Since data do not provide causal information, we MUST use models to assign causal effects to actions. It is also possible for models to incorporate information about goals of agents. But models have varying degrees of reliability, depending on how closely they are matched to the real-world causal mechanisms. Causal Effects have varying degrees of probable connection to action. It is hard to differentiate between a poor causal model and a weak causal effect. In any case, reality is so complex that our models are never exact replicas. Also, bad Models can successfully predict causal effects within a slowly changing environment. In this kind of a setup, correlations can substitute for causations in the short run. Correlations will lead to good forecasts and policy decisions in the short run. However, when the underlying structure changes, the

correlations will break down, and the bad models will fail. Something like this happened after the oil crisis caused by the Yom-e-Kippur War of the early 1970s. There was a structural change in the underlying economic realities, and most macroeconomic models broke down, producing wildly wrong forecasts. Later research showed that good models can match reality over a broader range of environments.

Causality for ACTION

The difficult work of exploring hidden causal structures is usually done with the intention of improving outcomes. We learn about causal mechanisms in order to be able to act more effectively to achieve our goals. When we ask the question “Does Engineering discriminate against women?”, it is because we are interested in taking actions to remove disparity and to end discrimination. To make effective policy decisions, it is essential to know the reason why there are so few females in Engineering. Consider, for example, the following two possible explanations.

- The male chauvinists want to keep females out of the profession.
- Admissions is gender-blind but uses SAT Math Scores as the basis for admission.

The Intervention Strategies are VERY DIFFERENT in the two cases. In the first case, we might try to make laws, use persuasion, or add female members to the admissions committee. In the second case, we might want to work on coaching female candidates on how to improve math scores, or persuade the university to provide remedial math courses to deficient candidates, while trying to balance gender.

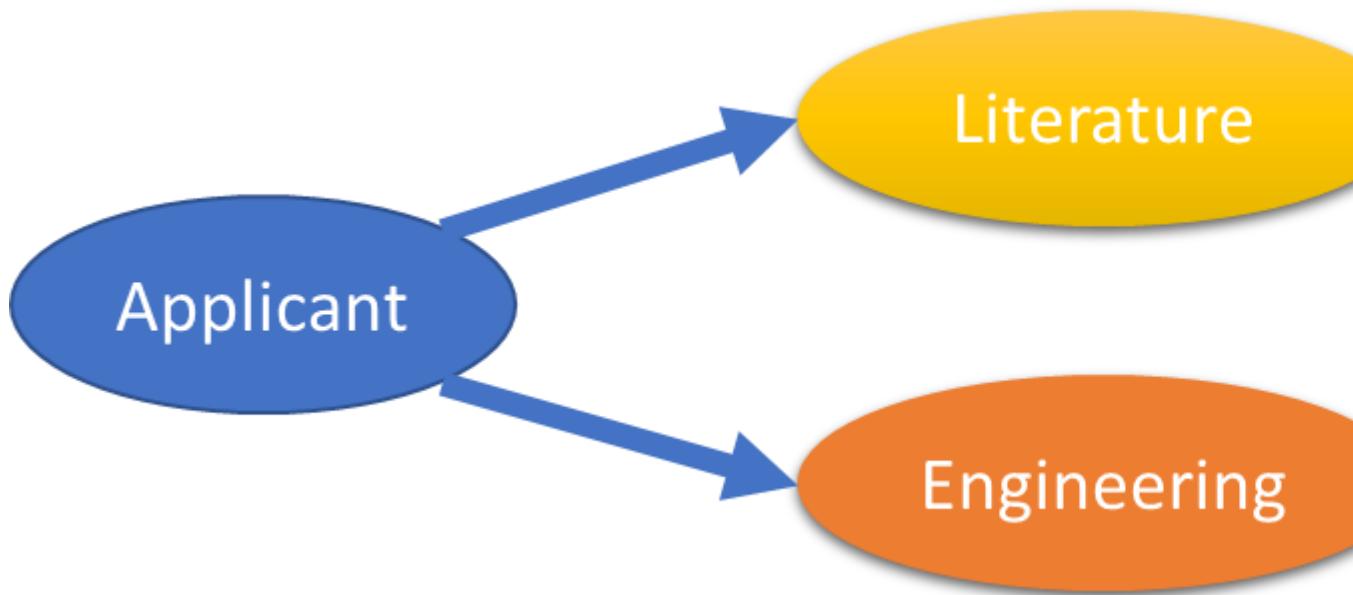
Agent-based causal effects simply cannot be learned from the data, since they are based on the intentions of agents. To understand the intentions behind monetary policy decisions, the minutes of the meetings are very useful. This approach is aligned with our general framework of “Causality as Child’s Play”. Experiments show that babies are good at reading intentions. ALSO, when babies see objects move, they look for agents, who caused that motion. They are aware that objects are not agents, and cannot move on their own.

Concluding Remarks

Causal Models are based on attempting to discover if action X causes consequence Y ($X \Rightarrow Y$). Woodward provided an improved account in “Making Things Happen” by asking the question “Can I manipulate X to create changes in Y?”. But this fails to take into account the GOALS of the Agent: How can I ACT to create DESIRED consequences? The Goals and Desires are Unobservable. But failure to achieve a Goal is one of the central building blocks of human experience. EXPERIENCE is the name of learning from failure and modifying actions to try to achieve the desired outcomes. Unobservables are central to causality. Also, human agency is central to causality, according to this account. Both are held in low-esteem by philosophers, which is why confusion over causality has lasted for centuries.

CAUSAL PATH DIAGRAMS

In the causal sequence $X \Rightarrow Y \Rightarrow Z$, Y is called a MEDIATOR. The Admissions paradox is described as Gender \Rightarrow Dept \Rightarrow Admit Ratio, where Dept is a mediating variable. It seems much clearer to diagram it as:



Even if both departments offer gender-blind admissions, if more females apply to the BIG department, and more males to the SMALL department, university-level admissions will show a bias towards females. This bias is due to females choosing the easier admit department, males choosing the difficult one, but is reflected in biased admit ratios at the university level.

12B Who is the Better Batter?

The goal of this lecture is to show that the apparently very simple question of the title has a very complex answer. In particular, data about match scores for two players cannot be used to answer the question. Questions about the causal relationship between players skills and scores come into the picture. Answers to these questions depend on deep knowledge about the real world aspects of cricket matches, going far beyond the numbers. Furthermore, the question about who is better cannot be answered without knowing the purpose of the comparison.

Another Example of Simpson's Paradox

We start by presenting another example of the Simpson's Paradox, based on cricket scores. All data below is artificial, with imaginary cricketers; however, we choose real names of cricketers to add some spice to the discussion, and make it come alive in minds of cricket fans.

We want to look at the question: “Sangakhara (S) and Miandad (M) are two batters. Which one is better?”. The data given to us is the following:

Over his career, S scored an average of 60 runs per match. M scored an average of 50 runs per match.

So, on the face of it, the numbers seem to prove that S is the better batter of the two. But do they? As already discussed, numbers are not capable of proving anything. We must have specific and detailed information about real world structures which generate the data, and also about the purpose for which the comparison is desired. To show this, we add a further detail about real-world cricket.

There are two types of Bowlers: Pace & Spin. Pace bowlers rely primarily on speed, while Spin bowlers try to spin the bowl to create unpredictable trajectories. Now assume the following:

1. M does well against P-type bowlers: Average 80
2. M does poorly against S-type: Average 30
3. S also does well against P-type: Average 70
4. S does poorly against S-type: Average 20

According to these numbers, Miandad is better than Sangakhara against P-type and ALSO against S-type bowlers. Even though he is better against both types of bowlers, he could end up with a worse overall score. This will happen if M faces many more Spin-type bowlers, which result in low averages for him, while S faces many more Pace-type bowlers which result in high average scores for him. For example,

- M played 100 matches: 40 against P-type, 60 against S-type
- M Average = $0.4 \times 80 + 0.6 \times 30 = 50$
- S played 100 matches: 80 against P-type, 20 against S-type
- S Average = $0.8 \times 70 + 0.2 \times 20 = 60$

Overall Average for S is better because he had a favorable environment, not because he is more skilled. This is a Simpson’s paradox because the overall average shows that S is superior, but when we break the scores down into two subgroups – batting against P-bowlers and against S-bowlers – that M is superior in BOTH subgroups.

NOTE ON REALISM: In any batting match, batters face both types of bowlers, and so this story is radically oversimplified. We can improve the realism while creating the same types of statistical issues, but then the numbers get complicated and impede understanding. So we have chosen a drastically over-simplified narrative for pedagogical purposes.

Who is better batter: M or S?

This analysis does not provide us with an answer to our original question. Perhaps surprisingly, the original question cannot be answered unless we know the PURPOSE for which a comparison between the two is desired. We must ask “WHY are you asking this question?” GOALS matter for the answer.

Purpose 1: Which of them should teach the other how to bat better?

M is better than S against both P-type and S-type. Therefore, M should teach S how to bat.

Purpose 2: Given a fixed set of bowlers (exogenous), which cricketer should be sent to bat?

Answer M is better than S against both types, so again M is better.

Complication: If the choice of Pace and Spin bowlers by the opposing coach is endogenous (depends on the batter) than the above answer may be wrong. For example, suppose Opposing Coach is AWARE of weakness of Miandad against S-type bowlers, but is NOT AWARE of same weakness of Sangakkara. In this case, Miandad will mostly face Spin bowlers and do badly, while Sangakkara will do well against a normal field of bowlers. It is also possible that the Coach is aware of the weakness of Sangakkara, but not of Miandad, in which case Miandad will do better.

ENDOGENEITY: Choice of batter DETERMINES or CAUSES composition of S- and P- type bowlers.

Alternative Configurations

Different types of configurations of skills for the two players lead to different conclusions. For example, suppose that:

- MGOOD against P-type: Average 80
- MBAD against S-type: Average 30
- SBAD against P-type: Average 20
- SGOOD against S-type: Average 90

Depending on percentages of S & P, M can have overall average between 30 and 80. S can have average between 20 and 90. Thus, we can see that higher overall score does not tell us ANYTHING about the skills of the two batters. In order to make comparisons, we must know WHY – more specifically, what action is to be taken on the basis of the comparison.

Purpose of Comparison?

Consider the issue of “Who can teach the other?”. Both players have different skills. So, M should teach S for P-type bowlers. S should teach M for S-type bowlers. There is no single answer to this question.

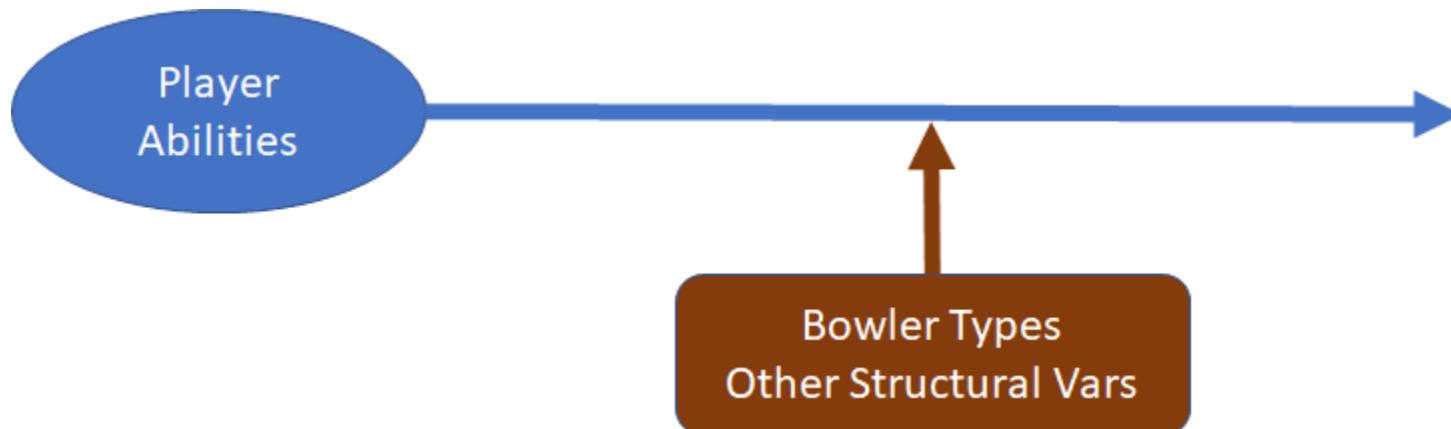
Who should we send out to bat? Against a fixed, exogenous, percentage of P- and S-type: M is better if P-type are dominant, and S is better if S-type are dominant.

If opposing coach is aware of both M and S, he will choose the worst type, so M will get 30 and S will get 20. In this case M is better.

We can multiply scenarios, and the choice of who is the better batter will vary with very specific details of a complex reality. This shows that causal information cannot be captured by the data, and also that multidimensional skills cannot be reduced to a single number for purpose of comparison.

Moderating Variables

From the specific example, we come back to theory. A moderating variable effects the strength of causal relationship, without directly being part of the causal sequence.



Many kinds of variables can effect the causal relationship between player skills and the scores. We have focused on Bowler Types, but there are other important factors like the nature of the pitch, weather, home versus foreign arena, etc. These factors do not enter the causal chain, but affect the strength of the causal relationship. In the previous lecture, we studied “Mediating Variables”, which actually enter the causal chain. We will review this concept below, for better contrast and understanding.

Mediating Variables

In the (artificial) Berkeley example of the last lecture, gender determines choice of department (Engineering or Literature), and the departments have their admissions policies which determines the admit ratio. The causal path diagram is as follows:



Suppose that both departments have gender-blind admissions, but Engineering is very choosy and has a very low admit ratio of 10%, while Literature is easy to get into, with an admit ratio of 90%. Also suppose the females overwhelming go for Literature, while males overwhelming go for Engineering. Then the overall admit ratio for the university will show a heavy bias towards females, even though both departments have gender-blind admissions. Again, this is a form of the Simpson's paradox.

A key factor which differentiates mediation from moderation is “screening”. In the above diagram, the mediating variable “Department” screens off the effects of gender on admissions. This is a technical term which indicates that gender influences choice of department, but once we know the department, that by itself is sufficient to tell us the admit ratio – we no longer need information about gender, because that information has already been captured by its effect on the department. Partial moderation occurs when some of the effect of gender is screened by the department, but there still remains a direct effect of gender on admissions. Such a case will be studied in greater detail in the next lecture.

Concluding Remarks

Perhaps the central lesson here is that the simple question “Which batter is better?” does not have an answer. Surprisingly, the data on performance is not enough to answer this question. In fact, when there are multidimensional skills involved, there may be no right answer. One batter may be better in one dimension while the other one dominates in some other dimension. Sometimes the answer is to be the basis of an action, a choice to be made between the two. In such a case, the question of “WHY do we want to compare the two?” must be answered. Different goals for comparison can lead to different answers as to which is the better batter. Very subtle and difficult to determine aspects of reality can determine which of the two choices is better in any particular choice situation. In the particular example chosen, we need to know how much the opposing coach knows about the two players, and how he will act in response to them. These factors cannot be determined with any degree of precision, and certainly no mechanical analysis of the numbers can find answers.

All of the analysis of this example is based on a vastly over-simplified ASSUMED underlying structure of reality: Different performances against S- and P- type bowlers. This structure is NOT present in the numbers, but only in the underlying reality. The real-world structure could be far more complex. An expert who spends a lifetime studying cricket would have far more knowledge of the factors which affect performance of players. Other HIDDEN variables may have important impact on causal chains. We have no way of knowing this from the numbers alone. This is why “big data” can never substitute for expert knowledge of ground realities.

Some Meta-Observations

What has been discussed is simple and obvious – if batters have different types of skills, then there is no objective way to merge the skills to come up with a single score for objective comparison of the two. A specific purpose (like choosing who to send into the field) and a knowledge of the specific conditions (whether P-type or S-types dominate, and how the opposing

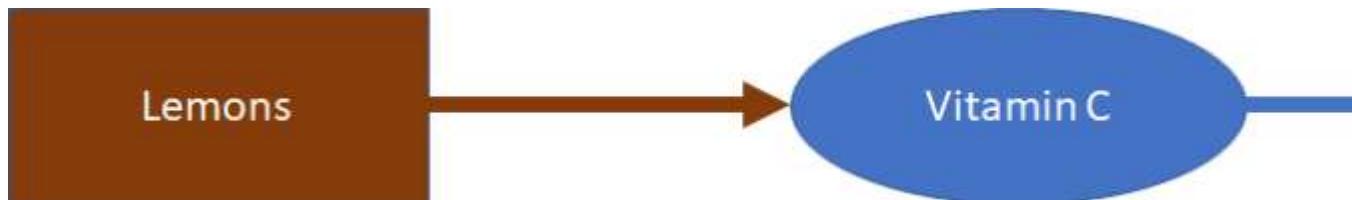
coach behaves) is required to make a good choice. Even with a one-dimensional score, we cannot objectively find out who is better. When it comes to multidimensional variables like intelligence, progress, research quality, college rankings, the problem of comparisons becomes even more complex. A single number CANNOT be used to evaluate quality, without use of subjective value judgements. Even in such situations, relevant environmental variables may not be known, making an informed choice impossible.

Even though these lessons are simple and obvious, the opposite is widely believed. Most of the population is firmly convinced that there is an objective way to determine if one batter is better than another, if one student is superior to another, if one university has a higher ranking than another. Why is this illusion so widespread, and becomes the basis of outrageously flawed policies? This story is very complex, and some pieces of it have been sketched in earlier lectures of this course.

The splitting of Christianity into Catholic and Protestant factions led to centuries of brutal and ruthless warfare, leading Europeans to reject religion altogether. With loss of faith in God, they developed the philosophy of empiricism which takes observations of external reality as the sole basis for knowledge. There is no question that our human experience is the sole basis for knowledge, but Europeans rejected our internal psychological experiences as being prone to error – since they could testify to the presence of God (who does not exist). This strong focus on external observables and neglect of internal qualitative human experience led to a very lopsided basis for knowledge. Increasing knowledge of the external world was combined with decreasing knowledge of our internal human emotional and spiritual world. The idea that knowledge is only about observables, and this can always be quantified led to Lord Kelvin's Blunder: Knowledge is always QUANTIFIABLE. This has led to the effort to measure many characteristics previously considered as unmeasurable and qualitative, and a rise in the prestige of statistics. Many widely used measures – such as IQ scores, SAT scores, College rankings, GNP – are MEANINGLESS NOISE but widely accepted by all and used to make many types of absurd arguments. This is the truth behind the aphorism that there are “Lies, Damned Lies, and Statistics”.

12C Partial Mediators & Simpson's Paradox

The previous lecture ([Who Is the Better Batter?](#)) discussed mediators: these lie on the causal pathway between a cause and its effect. For example, Vitamin C mediates the effects of lemons on scurvy:

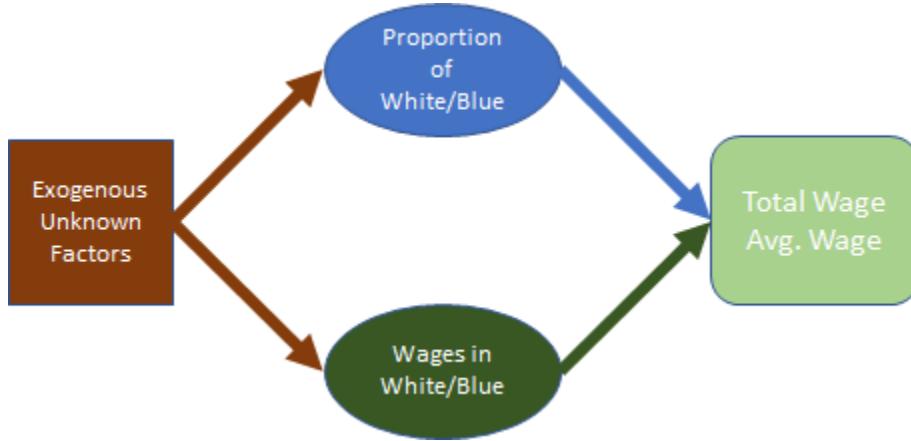


F12C.1 Lemons => Vitamin C => Scurvy

Causal path diagrams of partial mediation are used to provide a simple explanation of the Simpson's Paradox. The cause leads to a direct effect and an indirect effect via the mediator on the target variable. If the two effects oppose each other, the total effect can be the opposite of the

direct effect. This mechanism is used to illustrate and explain Simpson's paradox within the context of multiple real world examples.

The above diagram is an example of full moderation. Partial moderation occurs when there are other channels from the cause to the effect. We start with a recent example of Simpson's paradox in US Wages. It turns out that the overall US wages have been increasing, even though when we look at each educational subgroup, wages have been falling. How can wages fall in all educational subgroups, but rise for the labor forces as a whole? This is because the economic process acts through time on wages via two different channels:



F12C.2 Two

Channels to Average Wage

With time, wages in each sector (simplified to two classes) has been going down, exerting a downward effect on average wages. However, the proportion of educated (white collar) workers has been increasing, which exerts an upward influence on average wage. The combined effect can go in either direction. When the combined of the two channels differs from the partial effect of one of the channels, we get a Simpson's paradox. The paradox arises because we do not take into account other channels by which the cause can affect the outcome. A specific numerical examples is provided in the slides/video as a concrete illustration of how this would work.

The Berkeley Admissions Paradox Revisited

Exactly the same analysis applies to the Berkeley Admissions Paradox discussed earlier.

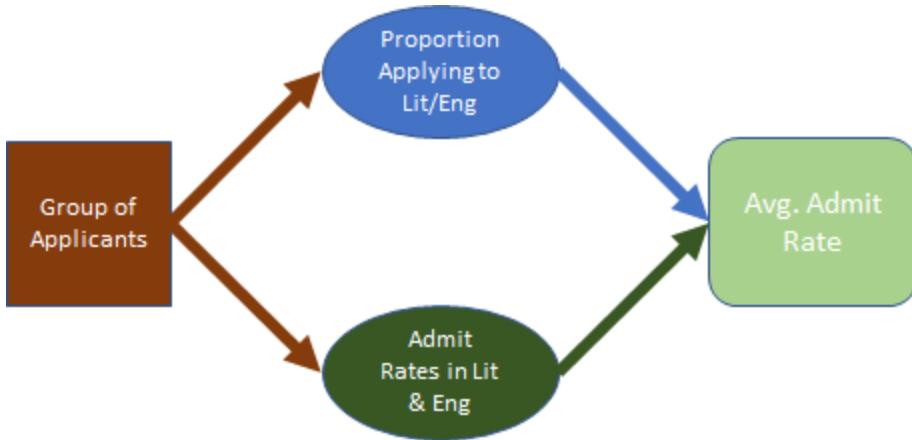


Fig 12C.3 Two

Channels Affect Average Admit Rate

Suppose Lit(erature) admit rates are around 20% and Eng(ineering) admit rates are around 50%. Suppose a group (females) receives preferential treatment in BOTH departments: 30% and 60% admit rates. The Dept. channel will create overall POSITIVE effect on average admissions for females.

BUT there is another channel which affects overall average. If females chooses mostly to apply to Lit, they will achieve 30% average admission overall. If Males apply mostly to Eng, they will achieve 50% admit rate, higher than the 30% of females. If females had CHOSEN to apply to Eng, they would have achieved 60% because they are favored. But they did not CHOOSE to do so. This shows how counterfactual analysis is closely related to causality — we need to contemplate what might have happened, to get a better understanding of the causal connections.

When this problem actually came up, statistician Peter Bickel analyzed it in the manner discussed above, and distinguished carefully between “bias” and discrimination”. Bias was a statistical phenomenon of observing a greater percentage of admitted males. Discrimination was an active effort by Berkeley to admit more males or females. Bickel observed that all the departments were actually attempting to rectify the bias by favoring female applicants. But the overall admissions process, driven by departmental choices made by females and males, led to a bias in the overall admissions rates against females.

The BBG (bad/bad/good) Drug

This is another example of the Simpson’s Paradox, but one with an extra feature. Depending on the causal relationships, which are not part of the data, either the overall ratio, or the subgroup ratio, can be the correct measure for comparison of performance of the treatment.

To assess the performance of a drug in curing a disease, we compare the recovery rates in two groups. The Treatment Group of 1000 people is given the drug, and 56% of them recover. The Control Group of 1000 people do not receive any treatment; only 44% of them recover. It appears that the drug is good, and increases recovery rates by 12%, However, we subdivide the patients into two populations: HBP and NBP for High and Normal Blood Pressure respectively.

When we look at these subpopulations, the drug appears to reduce recovery rates – it is BAD for both subgroups!

Control Group: recovery = 440

- 100 HBP: => 80% rec = 80
- 900 NBP: => 40% rec = 360

Treatment Group: rec = 560

- 900 HBP: => 60% rec = 540
- 100 NBP: => 20% rec = 20

As the data shows, in each group — HBP and NBP — recovery rates are lower with treatment than without treatment. But the overall recovery rate with treatment is higher. Why is this? How can it be? Again, it is a case of two separate mediating factors, which act in opposite directions. HBP group has higher recovery rates – with or without the drug. So any group which has higher percentage of HBP patients will have higher recovery rates.

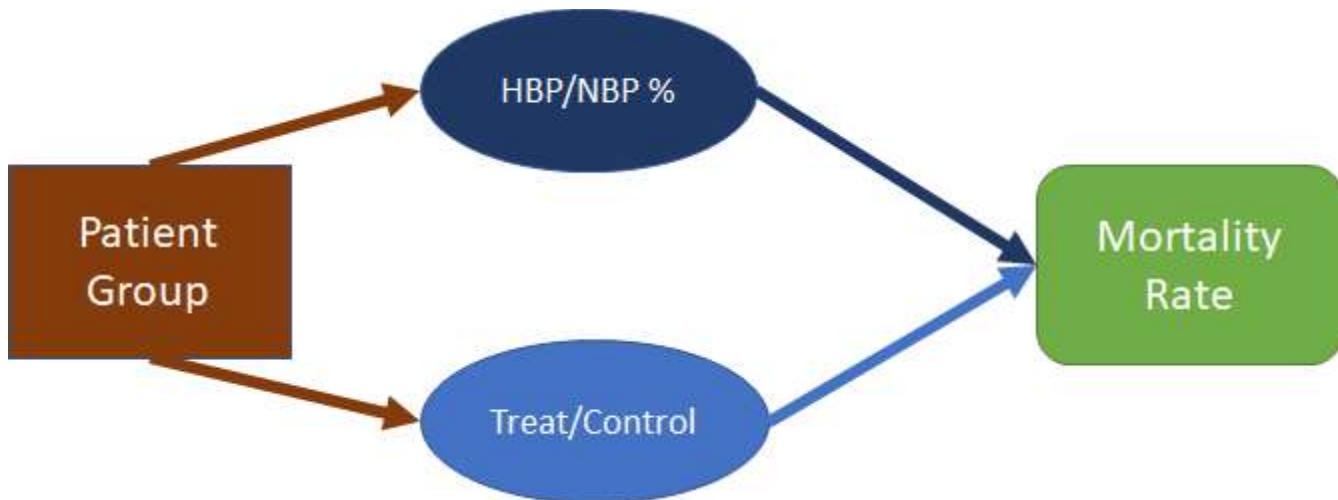


Fig 12C.4 The BBG Drug

The Recovery Rate in a group of patients is affected by two factors. The Drug is a NEGATIVE factor; it lowers recovery rates by 20%. But HBP is a positive factor, It raises recovery rates by 40% over NBP. If the treatment group has a very high percentage of HBP patients, then the positive effect of the HBP will overwhelm the negative effect of the drug and lead to a reduced mortality (or increased recovery) rate.

Now the MILLION dollar question is: WHICH of the two figures should we trust? Should we go by the group averages and say that the drug has good overall performance, and so we should give it to patients suffering from this disease. Or should we trust the subdivision into HBP and NBP, and not give the drug to anyone, since it is harmful to both types?

The answer depends on the causal relationships which can only be discovered by examining real world mechanisms, going beyond the data available. There are two possibilities.

The HBP is exogenous – that is, it is not affected by the partition of the group into treatment and control. For some reason, patients in the treatment group were chosen largely from the HBP population, while patients in the control group were chosen largely from the NBP population. In this case, the HBP/NBP proportion is a CONFOUNDER. It is strongly correlated with Treatment/Control partition, and it also affects the Outcome. The simplest possible path diagram for a confounder is as follows:

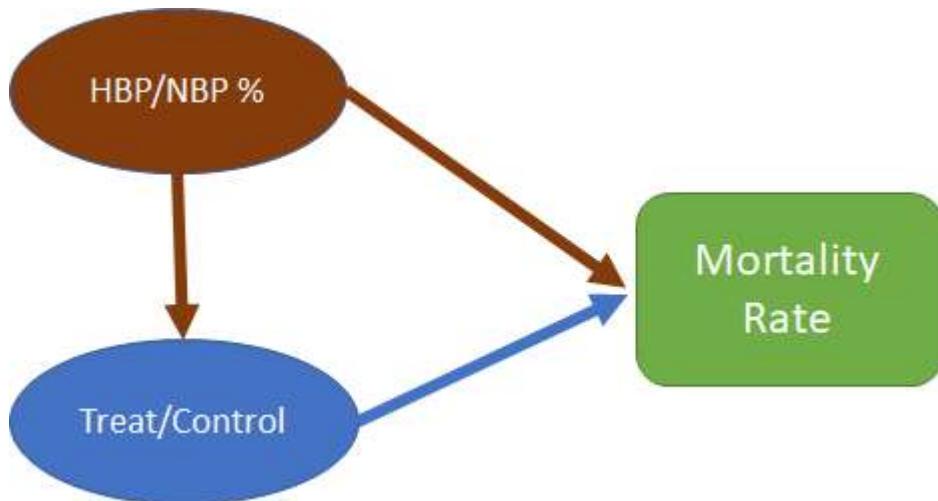
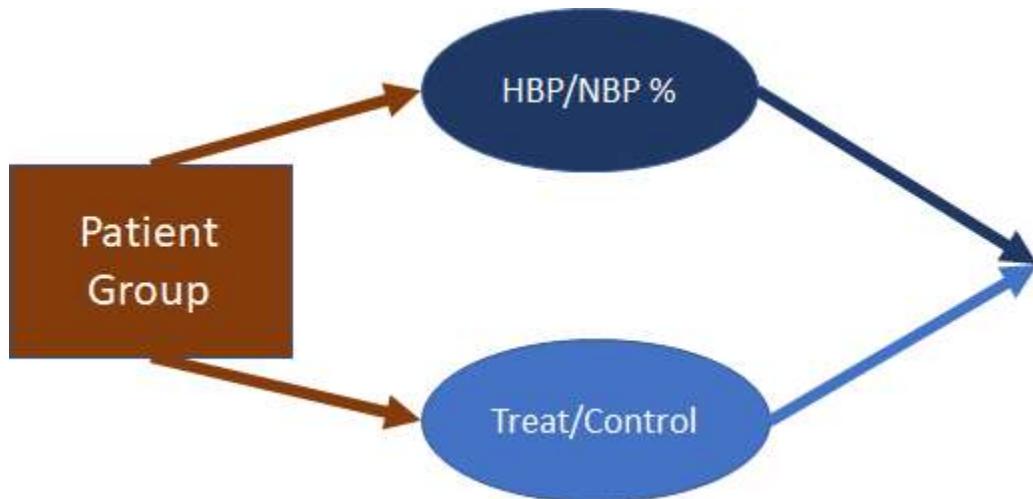


Fig 12C.6

Simple Confounder

If HBP causally effects the separation into Treatment and Control Groups, and also affects the Mortality (or Recovery) rate, then it is a simple confounder. This can happen if the group self-selects into Treatment or Control. Subjects of the study CHOOSE whether they want to be in the treatment or the control group. For some unknown reason, HBP people mostly choose treatment and NBP people mostly choose to be in the control group. There are situations in which such a narrative can be plausible, but the present case is not one of them. Instead, this is a more complex case of confounding, where the confounder is correlated with the cause, but does not cause it:



Fig

12C.7 Correlation due to Common Cause

In order for HBP/NBP ratio to create a problem in the analysis of effect of Treatment vs Control, it is enough for the ratio to have a correlation. In this case, if we look at relationship between Treatment and Outcomes, ignoring the variation of HBP/NBP%, we will attribute effects of the ignored factor to the Treatment. Some solutions to the confounding problem based on conditioning have been discussed earlier. In the next lecture, we will discuss the original solution created by Sir Ronald Fisher: randomization. In the present lecture, we want to look at another possible causal path, which completely changes the analysis and the conclusions.

The status of the variable HBP/NBP ratio, one of the determinants of the average mortality rate within a group, determines whether the subgroup analysis is valid, or whether the overall group average is valid. There are two cases. One is the HBP/NBP is EXOGENOUS. This means that there is no causal path from Treatment/Control variable to HBP/NBP. The other case is that HBP/NBP is ENDOGENOUS. This happens when there is a causal path from Treatment/Control to HBP/NBP. What this means in plain language is the the Drug Treatment CAUSES High Blood Pressure (HBP). In this case, the causal path diagram looks like this:

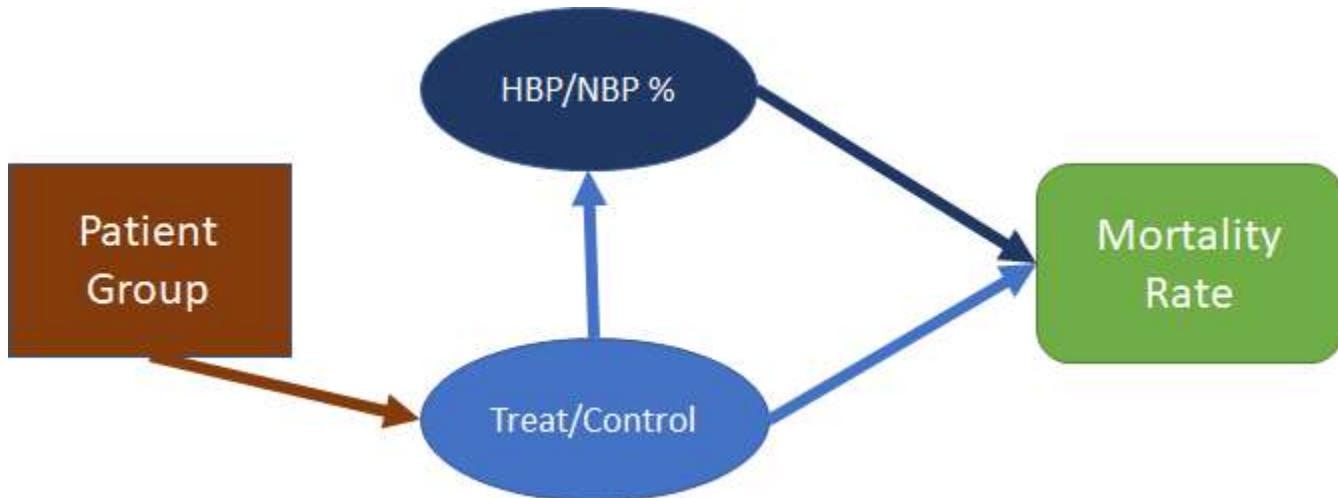


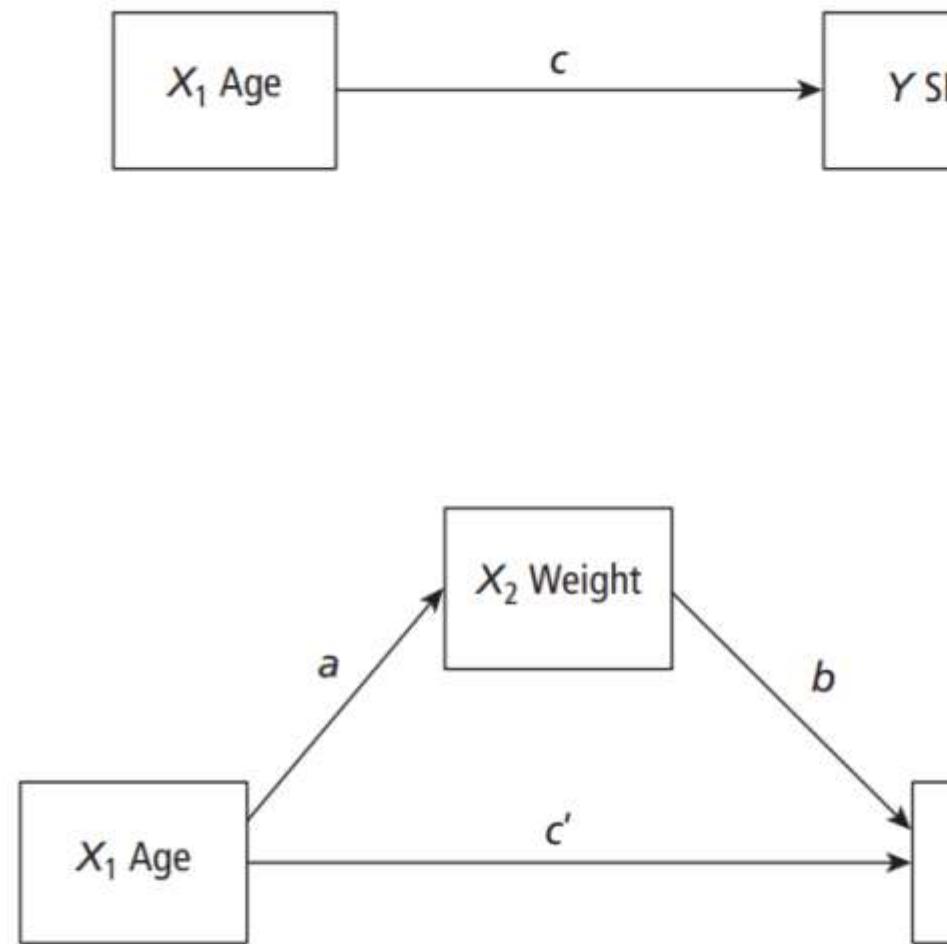
Fig 12C.8 HBP/NBP as Partial Mediator

If the drug itself causes High Blood Pressure, that explains why there are so many HBP patients in the treatment group. In this case, the causal path diagram has an interesting and simple explanation. The drug itself is harmful and hurts chance of recovery for both types of patients: HBP and NBP. However, one of the strong side effects of the drug is that it raises blood pressure — a population with 90% NBP and 10% HBP changes to 90% HBP by the effect of the drug. Now it turns out the HBP somehow substantially increases chances of recovery. So the side-effect of the drug creates an indirect pathway which is helpful in curing the disease. If this narrative is really the correct explanation, then one could improve outcomes by trying raise blood pressure in other ways, which not have the adverse effect that this drug has. HBP by itself creates a cure rate of 80%, but the drug lowers this to 60%. If we could raise HBP without this harmful side-effect, then we could achieve cure rates of 80%. Of course, all this is speculative on the basis of numbers, and real investigation would need to done in the clinics with the patients to see which causal pathways are valid.

This example also shows how the Simpson's Paradox operates: one cause has two pathways to the effect, one of which is positive and the other is negative. The total effect – the sum of the direct effect and the indirect effect – can go either way and conflict with the direct effect. This example also shows that causal paths are not in the data, but must be learned by looking at real world mechanisms involving investigations of the effects of the drug on patients.

Wrapup & Conclusions

The central concept under study in this lecture is that of partial mediation. If $X \Rightarrow M$ and $M \Rightarrow Y$, then M is a mediator on causal path between X and Y. If M is full mediator, M screens the effect of X on Y. Once we know M, we can predict Y, without knowing X. M is partial mediator if X acts on Y through other channels. Simplest case is DIRECT causation: $X \Rightarrow Y$. In this case, total effect of X on Y is the sum of the two effects, DIRECT and INDIRECT via M. For some policy questions, it is useful to be able to separate the two effects. Various techniques are in use to analyze mediation, to assess whether or not a given variable is a full or partial mediator, and also to calculate the direct and indirect effects. We give two more real world examples of partial mediation from the literature:

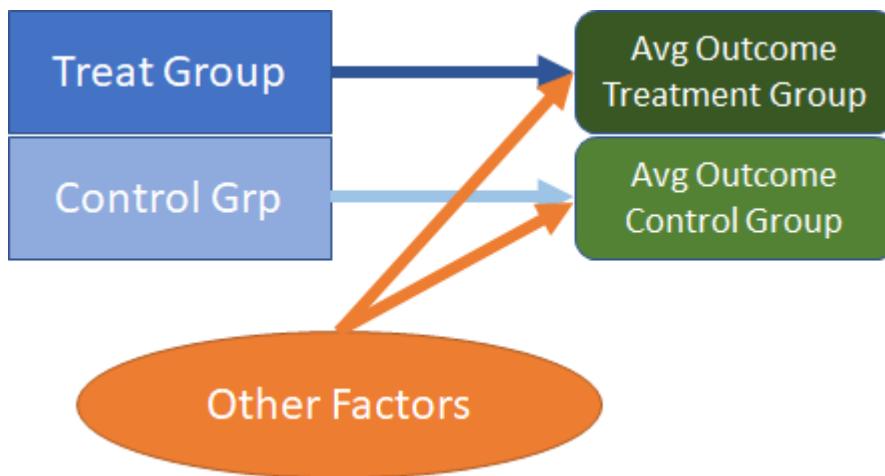
Figure 16.1 ♦ Hypothetical Mediation Example: Effects of Age on Systolic Blood Pressure

NOTES: Top panel: The total effect of age on SBP is denoted by c . Bottom panel: The path coefficients in this box represent the standardized path coefficients, $c = (a \times b) + c'$; the total relationship between age and SBP is the sum of the direct effect of age on SBP and the indirect or mediated effect of age on SBP through weight gain.

Age has a direct effect on blood pressure. Age also leads to increase in weight, which also effects blood pressure. If we can separate the direct and indirect effects then we can assess the value of weight control or diets in terms of their effect on the blood pressure.

12D Randomization as a Solution to Confounding

To evaluate the effects of a drug, we give it to a set of patients in a Treatment Group, and not to others in the Control Group. The difference in average outcome of Treatment and Control Group is an estimate of the effect of the Treatment. Confounding occurs when the Treatment and Control groups are different on some OTHER factor, which causes a difference in the outcomes. For example, if the treatment group consists of healthier patients, and the Control Group has sicker patients, then what seems to be the good effect of the drug may actually be a result of the better health of the patients in the treatment group. How can we rule out the influence of other factors?



L12D.1

Confounding

There are several strategies for PREVENTING other factors from influencing the DIFFERENCE in average outcomes. For simplicity, suppose that there is only one factor which matters: Gender. Females respond very well to the drug, while males do not. Then if the treatment group consists of males while the control group consists of females, the drug will seem ineffective. The first method to handle this problem is:

CONDITIONING: Prevent the relevant Factor from Varying: Fix Gender and then analyze each subgroup separately. This will lead to Two Different Effects: Male/Female. This raises the question: Does female AVG response difference in THIS experimental group GENERALIZE beyond the group? Answer: Females in group must be representative of females in general population. Then the question arises: How can we ensure this – that our experimental group represents the larger population? Also, what difference can it make if they are not?

Matching or Balancing: A second strategy involves making sure that the MIX of females/males is same in treatment and control group. Experimenter can choose the mix and equalize it in both groups. However, this also leads to the question: Does mix in this experimental treatment and control group REFLECT mix in general population? A similar

question is the following. Suppose all females have different responses to treatment: $f(1), f(2), \dots, f(n)$, but clustered around a common average. Can there be accidental imbalances within treatment and control – all females of one type in treatment, and another type in population?

RANDOMIZATION provides a neat solution to issues of representation — making sure that the sample is SIMILAR to the population with respect to the relevant factors. We first discuss a TECHNICAL result, heuristically. This is the key to Randomization as a rough equalizer: Let F be an UNKNOWN factor present in some percentage p of the OVERALL population. Let $R=R(1), R(2), \dots, R(1000)$ be a random sample of 1000 people chosen from this BIG population. A Random Sample has the technical meaning that EVERY person in population should have EQUAL chances of being in this sample. Let N be the number of people in the random sample having factor F . Then $N/1000$ is close to p with high probability.

In simpler words, the representation of characteristic F in the random sample is close to proportion of F in entire population. This holds whether or not we know anything about F , or whether or not we can observe it. The similarity of random samples to parent populations from which they come is the basis for the Gold Standard for Experimental Studies.

The Gold Standard: Randomized Experiment: If both Treatment and Control Group are chosen as RANDOM SAMPLES from the population, then for any factor F – known or unknown – proportion of F in sample will be close to proportion of F in population.

This does require large samples. Treatment and Control Group will be matched with each other on all known and unknown factors which can affect the outcome. Average effect of treatment and control within random sample should match average effect on population. However, there is **CAUTION**: Average effect on population may be different in SUBGROUPS. That is, if we match the mix of males and females in our sample to the population, we will get the average effect of the treatment correctly, but the average effect may be very different for males and females. We will not get any clue about this from the experiment.

We will now study a number of real world cases where randomization was used to learn about the world. These are all taken from Freedman, Pisani, and Purvis textbook Statistics. Relevant reference materials for these studies can be found from <http://bit.ly/FPPstat>

The Polio Study: First Let Us take A Naïve Approach. Polio is an EPIDEMIC disease. Here are some (artificial) numbers about the number of cases per year in the USA:

1952: 60,000 1953: 20,000 1954: 80,000 1955: 45,000

Suppose we give newly discovered Salk Polio Vaccine to ALL children in 1956 – what would we learn from the outcome in 1956?

Case A: 10,000? Could be due to effectiveness of vaccine. BUT ALSO, it could just have been naturally low in 1956, just like it was in 1953

Case B: 90,000? This could mean that the Vaccine did not work. BUT perhaps the natural number of cases would have been 120,000 without it.

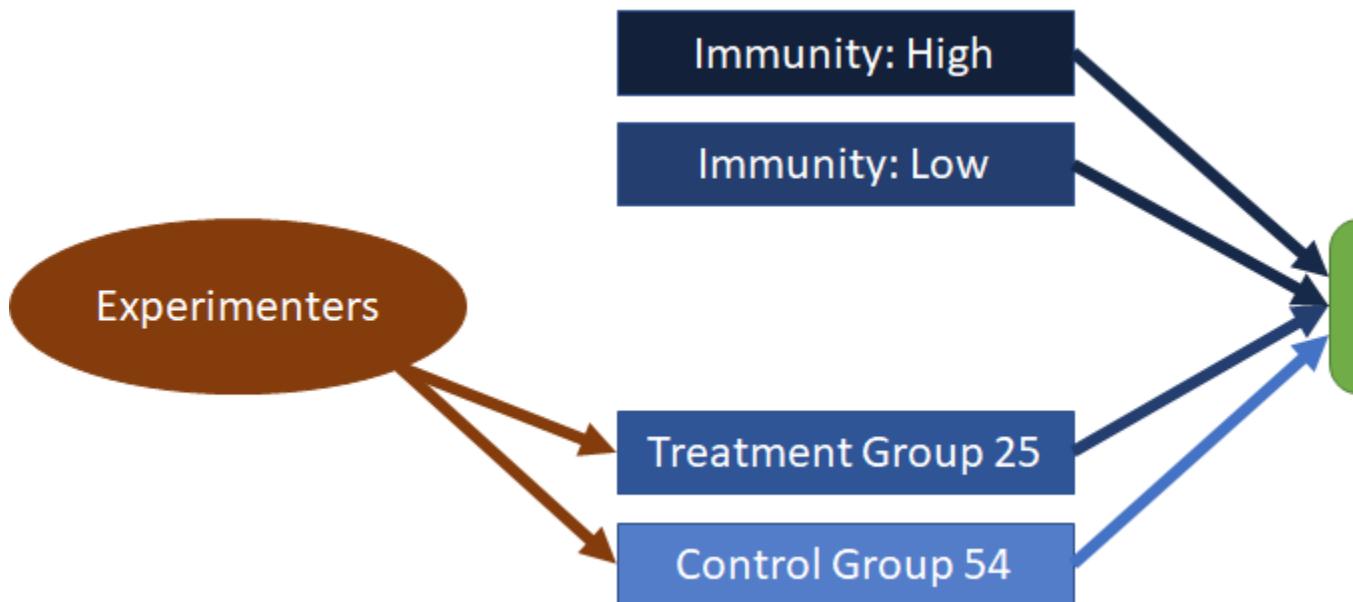
Nothing can be concluded in either case, because we have NO BASELINE for comparison? What WOULD have happened if NO ONE had received the Vaccine? To create this baseline, we always need to have a CONTROL group – similar to the Treatment Group, but without vaccine, so that we a basis for comparison.

The NFIP study: A large nationwide study was carried out with Grades 1 and 3 chosen as Controls, while Grade 2 chosen as Treatment Group. The idea was the children in grades 1 and 3 would be comparable to those in Grade 2. All unknown factors should be MATCHED across the groups. There were about 2M(illion) children chosen for experiment. 1M in Grade 2 as Treatment, and 1M in Grades 1 & 3 as controls. 0.5M refused consent in Grade 2. The outcome is given in the following table:

| Group | Number | Polio Cases | Rate/100,000 |
|----------------------|---------|-------------|--------------|
| Grade 2 (consent) | 225,000 | 56 | 25 |
| Grades 1 & 3 control | 725,000 | 392 | 54 |
| Grade 2: no-consent | 125,000 | 55 | 44 |

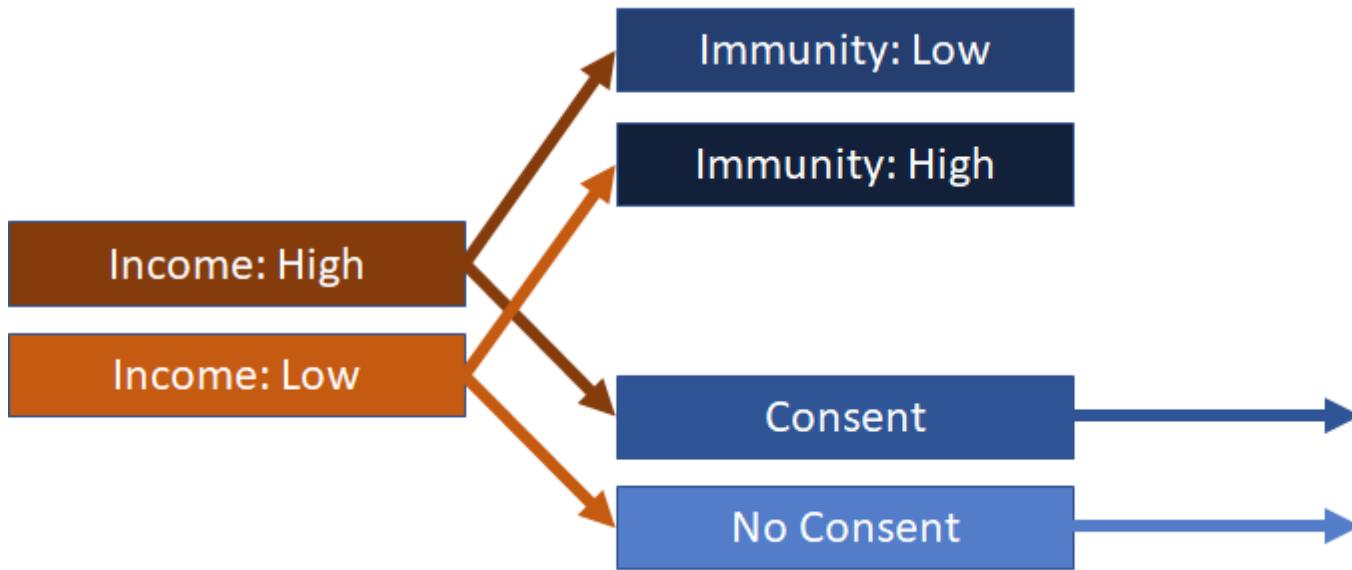
L12D.3 NFIP data

The control group had 54 per 100,000 cases while the treatment group had only 25 — this shows that the number of cases were halved by the vaccine, and shows that the vaccine was effective. Or does it? Can we be sure that the treatment group and the control group were similar with respect to relevant factors? One figure which suggests non-similarity is the polio rate of 44 per 100,000 in the Grade 2 no-consent group. This group did not receive vaccine, just like the Grades 1 & 3, but the polio rate of the two unvaccinated groups is quite different. WHY? It does not seem that failure to consent to vaccination should have any effect on susceptibility to polio.



L12C.4 Polio Immunities

We know that children have varying levels of immunity to polio. Do the treatment group and control group both have the same average level of immunity? The two groups are different with respect to one known factor: treatment group has only children whose parents consented to vaccination, while Grades 1 and 3 have all children, including those whose parents would not have consented. But how can failure to consent have any effect on polio rates?



L12C.5 Income Immunity

Upon investigation it turns out the parents who give consent have higher average income, while the no-consent group tends to have lower average income. But how does income affect polio? It turns out that high income families have hygienic environments with less exposure to germs and hence less immunity. The low income groups have higher exposure, and hence higher natural immunity. This explains why the polio cases are fewer in the no-consent Grade 2 group (45 per 100,000) as compared to Grades 1 & 3 controls, where the consenters and non-consenters are mixed. The common cause of Income creates a correlation between immunity levels and treatment/control which causes confounding. This creates a bias against the vaccine. One way to fix the bias would be to match incomes across the treatment and control groups. But this would only match the known factors approximately, and there might still be confounding due to unknown factors. Randomization takes care of these problems of matching relevant factors.

Randomization: For a randomized experiment, we first separate parents who consent from the non-consenters. The experiment will be done ONLY on the consent group, to allow for a match between control and treatment group. For each consenting child, we randomly assign them to either the treatment or the control group. This led to 200,000 children in the treatment group, and 200,000 in the control group. A remainder of 350,000 children were in the no-consent group. The outcomes of this randomized experiment were as follows:

| Groups | Number | Polio Cases | Rate |
|-----------|---------|-------------|------|
| Treatment | 200,000 | 56 | 28 |

| | | | |
|------------|---------|-----|----|
| Control | 200,000 | 142 | 71 |
| No-Consent | 350,000 | 196 | 56 |

L12C.6 Randomized Controlled Polio Experiment

There is a big difference between the 28 per 100,000 rate in the treatment group, and 71 per 100,000 in the control group. The vaccine has prevented about 40 cases per 100,000 of polio. Note the difference between the 56 rate in the no-consent group and the 71 rate in the control group. The lower rate in the no-consent group is due to the higher immunity levels in this group, as opposed to higher average income in the consent control group (which did not receive vaccine, even though they consented to receive it)

Double Blind: A subtle difference between the two groups is the fact of being the groups itself. A child who knows he is in the treatment group may recover from polio, or resist it, simply because of the psychological effect of knowing he/she is vaccinated. This is known as the “placebo” effect. To prevent this from happening, all children are given doses of “polio” vaccine, but the control group is just given flavored water, so that they do not know whether or not they have received the vaccine. Another bias factor comes from the doctors. Sometimes it is difficult to assess whether or not a child has polio. If the doctor knows that a child is vaccinated, the doctor might just him to be polio-free in ambiguous cases, and assess him to have polio in cases where he knows that the child is not vaccinated. To prevent this bias, doctors are not told which children have been vaccinated. This is called a double blind study, where neither the patient, nor the doctor, knows who is in the control group or the treatment group.

The Literary Digest Poll: In 1936, Democrat FDRoosevelt had completed his first term. Dealing with Great Depression, he supported deficit spending to help the unemployed. Running against him was Landon, Republican, who campaigned for balancing the budget. The Literary Digest (LD) predicted overwhelming victory for Landon, based on the largest poll ever conducted: 2.5 million readers, who responded to the poll. LD was very prestigious magazine and had successfully predicted election results since 1916. LD predicted the Roosevelt would get 43% of the vote (based on their non-random sample). He actually got 62%. This HUGE ERROR damage the prestige of LD, and Literary Digest went bankrupt soon afterwards.

Gallup's Random Samples: Around this time, the theory of random sampling had been developed by Fisher, and Gallup was attempting to apply it. Gallup took a “random sample” of about 50,000 people and predicted a victory for FDR. His prediction was 54% – actual vote was 62% – large error, but right result. Similarly, Gallup took a random sample of 3000 from Literary Digest readers, and predicted that LD would predict 44% vote for FDR. In fact, LD predicted 43%. This raises the following questions:

1. Why did LD make such a HUGE error?
2. How did Gallup make an accurate prediction of LD results?
3. Why did Gallup make a fairly big error on the OVERALL results?

Answers

LD's Huge Error: Readers of LD were NOT a random sample. Huge over-representation of Republicans – wealthier, more educated.

Gallup's Error: RANDOM SAMPLING is hugely difficult. One must ensure that EVERY VOTER has an equal chance of being in the POLL. There is no overall list of all voters – it is hard to predict who will actually go to the polls. However, methods have been improving and sample size have been going down. In recent times, Gallup uses samples of size 3000 to 6000 to predict the results for the US Elections.

Gallup's accuracy in predicting LD results: Full sample of Literary Digest readership was available, making it easy to create a random sample.

Surprising result: Accuracy of random sample doesn't depend on population size, only on size of random sample. Regardless of how large the parent population it, we can be 95% confident that a random sample of size 3000 will be accurate to within 2%. Actual accuracies (the discrepancy between the Gallup forecast and the actual election result) are generally a bit larger than this theoretical accuracy because of a number of practical problems not considered by theory.

Some Poor Methods for Choosing Controls

Even though random sampling is the gold standard, it is quite hard, sometimes impossible, to do. Many other alternatives exist and are often used in real world research. We go through some real world examples to show the weaknesses in these other methods for choosing controls.

NO CONTROLS: Pakistan has made great progress in improving infant mortality. The rate came down from 1960: 120 per 1000 to 2000: 60 per 1000. A 50% reduction over four decades is a tremendous achievement. TRUE/FALSE???

Answer: We cannot say. No benchmark has been provided. Some countries did better over the same period. Others did worse. Without any benchmark, it is impossible to say if this is good or bad. By CHOOSING benchmarks carefully, we can MAKE it appear good or bad. Using percentile-rank in all countries, Pakistan ranked in the bottom quartile throughout – neither improving nor decreasing in rank. So, it is an average performance, compared to other nations, neither good nor bad. But for different contexts, different benchmarks may be appropriate, and lead to different results.

Non-Random Controls: We reproduce Table 2 from Chapter 1 of Freedman, Pisani and Purves regarding the Portacaval Shunt Surgery. The table lists the outcomes of 51 different studies on this surgery. The well-designed studies show the surgery to have little or no value. The poorly-designed studies exaggerate the value of the surgery. Non-randomized studies come to the opposite, and wrong, conclusions, from those of the randomized studies.

| | | | |
|--------|-----------|-----------------|----------|
| Design | Very Good | Moderately Good | Not Good |
|--------|-----------|-----------------|----------|

| | | | |
|---------------------|----|---|---|
| No Controls | 24 | 7 | 1 |
| Non-random controls | 10 | 3 | 2 |
| Randomized Controls | 0 | 1 | 3 |

L12D.7 Studies of Portacaval Shunt Surgery

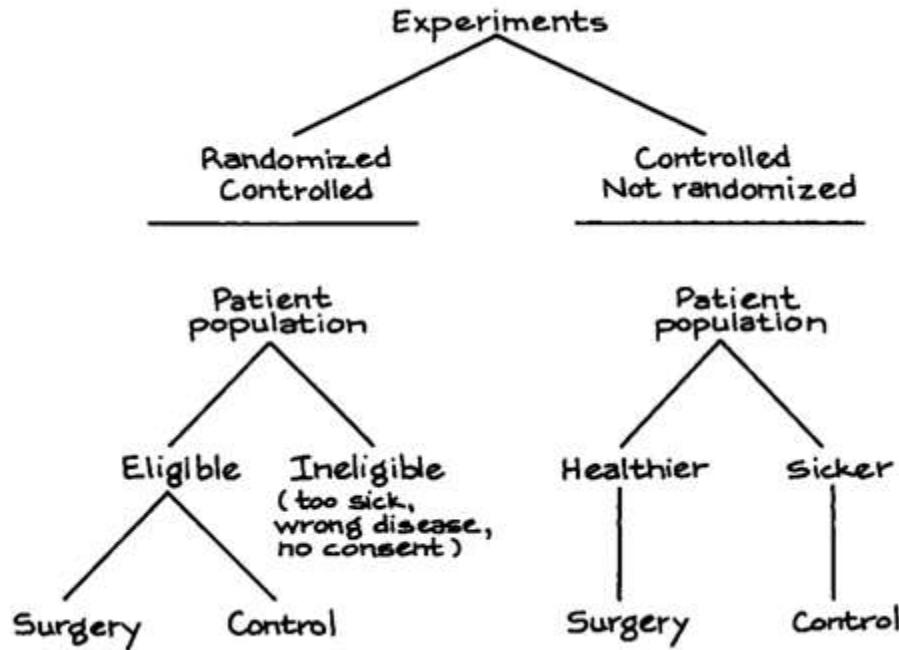
How do we know that the randomized controlled studies get the right result, while the uncontrolled or non-random controls are getting the wrong results? We can learn this by looking at the survival rates of the control groups in the two types of studies. These are given in the table for 3 year survival rates after Portacaval Shunt Surgery below:

| | Randomized Controls | Non-random controls |
|-----------------------|---------------------|---------------------|
| Treatment (Surgery) | 60% | 60% |
| Controls (No Surgery) | 60% | 45% |

L12D.8

In the randomized studies, both treatment and control group have the same survival rates of 60%, which shows that surgery does not provide any benefits. In the non-randomized control studies, the rate is the same 60%, showing that the surgery in these studies has the same result as surgery (or no surgery) in the randomized studies — comparing surgery to the randomized controls shows no benefits from surgery. **HOWEVER**, the controls in the non-randomized studies have a significantly lower 3 year survival rate of 45% — this explains why the non-randomized studies report good results: comparison with controls in poor health leads to the conclusion that the surgery is good, whereas the superiority of the Treatment comes from the better health of the Treatment Group.

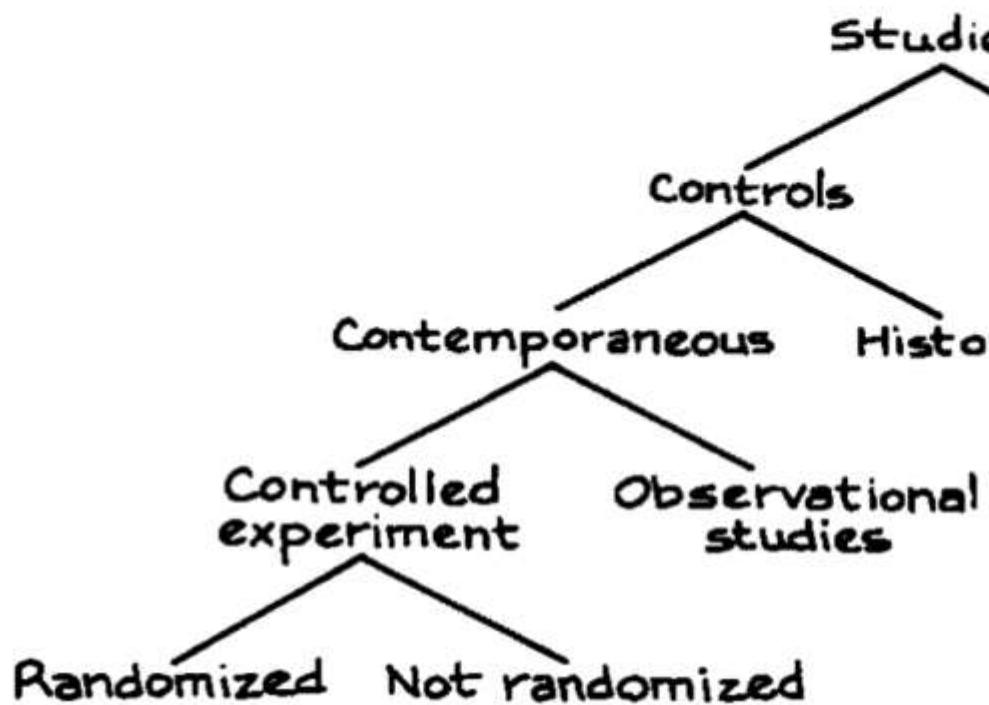
Why is the Treatment Group for Surgery generally better than the control group in non-randomized studies? The reason is that surgeons screen patients for eligibility for surgery — patients should not multiple complications, and other medical problems which can cause difficulties with the surgery. So patients chosen for surgery by doctors are generally substantially healthier than those who are not considered eligible for surgery. When these less healthy patients are chosen as controls, a bias naturally results. This selection process is described in the diagram, taken from Freedman, Pisani, and Purves, below:



L12C.9 Biased Choice of Controls in Surgery

The video lecture discusses a few other examples taken from Freedman, Pisani, and Purves, which show how non-random controls can lead to wrong results in experiments, and how these wrong results lead to wrong medical procedures being adopted, which cause a lot of damage. The discussion of this lecture can be summarized in the following diagram, which classifies the different types of experiments and studies made for comparing the outcomes of medical treatments:

4. When looking at a study, ask the following question: Was there a control group at all? Were historical controls used, or contemporaneous controls? How were subjects assigned to treatment—through a process controlled by the investigator (a controlled experiment), or a process outside the investigator (an observational study)? If a controlled experiment made using a chance mechanism (randomized controlled experiment), does it depend on the judgment of the investigator?



L12C.19 Different Kinds of Experimental Studies

Concluding Remarks:

An Exogenous Variable Changes, without being influenced by any of the variables within the system under study. FREE WILL creates exogenous variables. RANDOMIZATION also creates exogenous variables. When variables are exogenous, effects of change can easily be isolated – they are not confounded with other factors. Failure to randomize, and confounding, can lead to very serious errors. This is a common problem, with no easy remedy.

12E Mindless Data Crunching

Narratives and Data Analysis: Data makes sense in context of a narrative. Numbers, by themselves, do not talk. The same numbers lend themselves to different, even opposing

narratives. Six losses in wars can be woven into a narrative of “extra effort and courage is required for victory” or into a narrative of “failure is our fate”. The facts remain the same, but the narrative is dramatically different.

The same is true in statistics. Just think of a series of numbers – without a name, they are meaningless. But name them as “happiness”, or “IQ”, or “GNP”, and they suddenly acquire a life. Good narratives conform to ground realities represented by the numbers; these always go beyond the numbers.

Contemporary statistical methodology creates the illusion that data, by itself, is sufficient. It tells us that narratives are RHETORIC – biased opinions – which seek to shape the objective and unbiased data in favor of personal prejudices. This is the common understanding which drives statistical methodology today. In fact, Data are NEVER enough – It is ALWAYS a narrative which makes the data speak. Conventional Statistics brainwashes us into ignoring the narrative. For example, REGRESSION results presented in thousands of articles. Regressions provide results ONLY under assumptions never fully discussed and clarified to students. All of statistics operates in this way: the central importance of the real-world context for numbers is ignored.

Some examples as reminder: In the last lecture, we saw that the HBP as mediator causal chain (Treat => HBP => Cure) requires radically different analysis from HBP as confounder: Exp => HBP/NBP, Exp => Treat/Control (experimenter is common cause of HBP and Treatment) and Treat => Cure. Whether or not drug increases blood pressure is NOT in the data, and we can find which of the two is the correct causal path only by investigating the real world, not by data analysis. In general, given data series X and Y, it is impossible to link them via causal relationships, without a NARRATIVE which contextualizes and connects them. For example, when studying discrimination at Berkeley, we MUST talk to the admissions committee. Without such conversation, we would not know which data to analyze.

The idea that data must be interpreted within their real-world context to be meaningful seems trivially obvious. The puzzle is: “Why is the OPPOSITE idea dominant?” This requires some explanation, sketched very briefly below. For a more detailed discussion, see “The Emergence of Logical Positivism” (<http://bit.ly/AZelp>)

1. Rejection of Christianity led to trauma of loss of faith.
2. Empiricism was the response: Believe in only what you can touch and reason. REJECT the evidence of the heart.
3. Belief in observables and external world. Rejection of knowledge of our internal world.
4. Observables can be measured. Lord Kelvin’s Blunder: We have knowledge of something ONLY when we can measure it! See: <http://bit.ly/AZKelvin>
5. Increasingly meaningless measurements, completely missing the mark, came into vogue, and continue to be widely popular.

Because causality is always contained in the narratives about the data, and never in the data itself, major misconceptions from this positivist methodology surround the study of causality. This lecture discusses these statistical misconceptions and their harmful effects in greater detail.

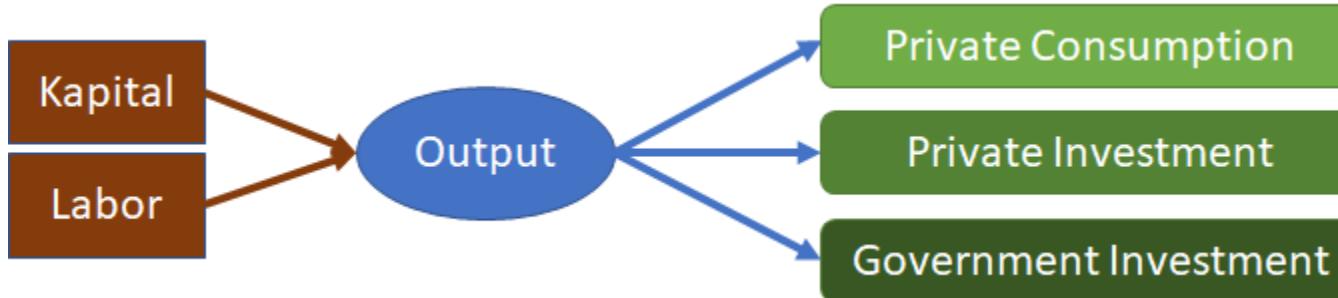
Exogeneity: Key to Causality

If event X occurs FREELY, without being caused by anything in the past, and THEN something else (Y) UNUSAL happens, we can GUESS that X caused Y. Philosophers ask: How would we know if it is VERIDICAL? (TRUE)? The answer is that “We don’t.” We must learn to live with imperfect knowledge and uncertainty. The misconception that only “certain” knowledge is worth having has been the source of an enormous amount of confusion in Western philosophy.

Our exercise of Free Will is the best way to generate exogenous events. We feel our own freedom to choose. The best way to find God is to pray to God, and FEEL the answer. Ever since Descartes, Western philosophers have discounted feelings as a source of knowledge, so one source of exogeneity is lost. ALSO, Newton’s laws led to belief in deterministic universe, losing RANDOMNESS as a source of Exogeneity. This failure of Western Philosophers to Understand Exogeneity has had Serious Consequences. We illustrate this by two examples:

Supply Side (SS) versus Demand Side (DS) Economics.

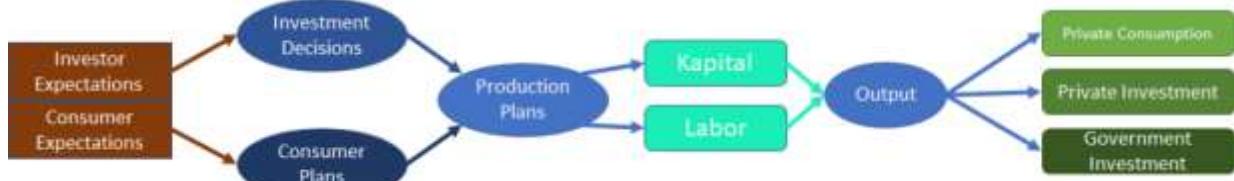
Supply Side is a causal theory about which factors are exogenous, and which factors are determined endogenously, within a capitalist economy. This is best understood via a causal path diagram:



12E.1 Supply Side Causal Path

SS: K, L are exogenously determined. Full Employment Holds. Output is $Q=F(K,L)$ – fixed exogenously. $Q=C+Inv(G)+Inv(P)$ If government invests (fiscal policy), other two factors will go down. Since output is determined by the input factors K and L, the idea of “crowding out” appears as a mathematical fact. If government investment increases, the other two factors composing the total output must decrease. Increase in demand of one sector can only cause a decrease in the share of others, or a rise in prices, or both.

Keynesian theory, which leads to Demand Side economics has a dramatically different view. It is the demands for products which lead to their production. The causal path diagram for demand side economics can be pictured as follows:



12E.2 Demand Side Causal Path

According to Keynes, Investment Expenditure (Demand) is the exogenous driving force for the economy. Investors invest according to their expectations about the future – these expectation are not anchored in reality (animal spirits). Investment decisions drive producers to hire suitable amounts of input factors Kapital and Labor required to meet the investment and consumption demanded. Demand (Investor+Consumer) $\Rightarrow Q \Rightarrow C$. Demand determines production plans and consumer income. This causal path diagram leads to entirely different (Keynesian) strategies for economic recovery and growth. The central point for us here is that it is IMPOSSIBLE to resolve these using DATA. We must study the real world to learn about what is exogenously determined!!

Supply & Demand Theory: Hopeless Confusion about Exo- and Endo- geneity

The traditional supply and demand story of economics is one of the central narratives of economics, and used to justify and explain a wide range of diverse phenomena. Nonetheless, it is hopelessly confused because of its failure to distinguish between exogenous and endogenous variables. The S&D theory posits that prices are endogenously determined via the equilibrium of supply and demand curves. However, the supply and demand curves are derived under the assumption that consumers and producers consider the prices to be exogenous. These curves are defined by :

- Demand Curve: Take price as EXOGENOUS, and ask: How much would you demand if the price of good G is P^* ?
- Supply Curve: How much would you supply if the price of the good G is EXOGENOUSLY fixed at P^* ?

When the price is endogenous, these questions do not make sense!! If a consumer is asked this question, he must be put into a state of ignorance about how the system works. What is the exact state of knowledge of the consumer about the system matters enormously but is left unspecified, leading to a number of paradoxes and contradictions within this theory. A fancy way to see these problems is to switch to a general equilibrium view of the economy, which looks at the system as a whole. In this system, all prices achieve equilibrium. It is impossible to CHANGE one price, and ask questions about what would happen, because that would throw the entire system out of whack. The questions about what happens in disequilibrium are not generally considered by economists, but must be an essential part of supply and demand curves as traditionally defined.

How Can We Decide Among Competing Narratives?

As we have seen, the data is unable to help us in this regard – the same data can be compatible with multiple narratives. The causal path diagrams are an essential tool for clarity in thinking. They provide clarity about the causal connections being asserted about the data. By themselves, they just articulate the causal assumptions clearly, but this does not help us to decide among them. However, the causal path diagrams do tell us the questions we need to ask, in order to distinguish between competing narratives. For example, how can we decide between the Supply Side and the Demand Side narrative? We need to ask:

- HOW do investors make investment decisions?
- How do consumers make consumption plans?
- How do producers make production decisions?
- How do price-setters make pricing decisions?

The two narratives offer different answers to these questions. Because economists are wedded to a positivist methodology, they do not believe in asking qualitative questions – they believe in the numbers. In his book, “*Asking About Prices*”, Alan Blinder has asked these questions, and found that the answers are in conflict with dominant narratives in economics. Positivist predilections in economics create a theoretical language which is not even able to formulate questions about narrative based causal relationships, let alone provide answers.

Brief History of Causality in Econometrics

This is a brief summary of an extended discussion in [Causal Relations via Econometrics](#): (<https://ssrn.com/abstract=1374208>). The Cowles Commission Approach gave us the first approach. Theory provides us with Exogenous Variables and Structural Equations (Equivalent to Causal Path Diagrams). Massive Predictive Failure of Macroeconomic models built on the basis of these method occurred following the Yom-Kippur War, Oil Embargo, and the resulting Stagflation. When exogenous shocks occur, the causal chains do not break BUT correlations break down under structural change. Failure to differentiate between causality and correlations had led to econometric models being built on strong correlations, most of which broke down due to external shocks. In the aftermath to this methodological crisis, three approaches were adopted:

1. Sims Critique: VAR models – Abandon all models, and go to purely NUMBER crunching – forget about real world context.
2. Lucas Critique: Deeper involvement of theory (But AXIOMATIC theory, based on capitalist ideology).
3. Hendry Critique: Continue to use correlations, but look for STABLE correlations – do much better and more sophisticated tests for stability of correlations.

Of the three, only Lucas looks for causal relations, but Lucas gets causes out of a FALSE theory, presented as axiomatic truth. When such causal assumptions conflict with empirical evidence, the empirical evidence is ignored by Lucas and followers; see “[Romer’s Trouble With Macro](#)“.

What is the right approach to forecast failure?

Causal chains persist in shocks, and models built on causal chains would have superior performance. But how can we discover causal pathways? We must start with the Cowles Commission approach: a theory based data analysis which specifies Exogenous Variables and Structure (causal paths) based on BOTH theory and empirical evidence. This is a starting point for causal discovery. The next step is to TEST the links in the causal path diagram. Given a causal path, there are many approaches to TEST for compatibility of correlations with a causal structure. The DEEPER problem is the DISCOVERY of causal structures. There is a common and widespread misconception that this can be done by BIG DATA. As we have seen, the narratives which provide causal linkages are not in the data.

Discovery of Causal Structure: Look for SHORT and STRONG, DIRECT causal paths. For example, it is a widely accepted theory that Pumping Money into an Economy will lead to INFLATION. To discover causal chains, ask HOW? The Central Bank lowers discount rate. Banks will increase borrowing from central bank. Overnight Loans from Central Bank (AND in interbank market) will rise.

- Study this FIRST round impact: How strong a response, how stable, and how it is affected by environmental and structural variables.
- THEN study SECOND round affects. WHAT do banks use money for? WHO do they lend to? What do lenders do with money?

TRACE the causal chain STEP-by-STEP – USING knowledge of the real world mechanisms. We can describe the process by saying that the search for causal mechanisms is the Search for MEDIATORS: if we know that $X \Rightarrow Y$, then we must ask HOW does X cause Y? What are the CHANNELS used to create the effect. In particular we look for mediators M which transmit the effects of changes in X to Y: $X \Rightarrow M \Rightarrow Y$

Also we must look at the INTENTIONS of the agents: What are the goals of the policy-makers? What are the models (explicit or implicit) they use to forecast effects of their policy actions? GOALS plus MODELS drive policies. We need to take both into consideration, to derive causal relations. This part is the DIFFICULT meta-analysis, involving DISCOVERY of causal structures.

Causal Path Diagrams

There are many advanced concepts in path diagrams, which have not been covered here. ONLY a simple exposition of basic concepts has been provided. ONCE we have a causal path diagram, subsequent data analysis is fairly simple. HOWEVER, this analysis is NOT reducible to regression models, or ANY of the standard statistical/econometric methods. WHY? Because these methods impose EXTRANEous assumptions which are generally not valid. This part – data analysis WITH an assumed causal path diagram – is relatively straightforward. DISCOVERY of causal paths is NOT simple, and cannot be done by data analysis. This involves looking at real world mechanisms, goals of policy makers, as well as the causal mechanisms in the minds of policy makers which lead to their actions.

Concluding Remarks

BIG Data and Machine Learning cannot go about learning about qualitative structures of the real world. Robert Shiller in Narrative Economics explains how conversations at dinner tables were of great importance in creating narratives which led to the Great Depression. “Virality” of such conversations had an impact on expectations of people. This is similar to Keynesian Economics which argues that People’s Expectations matter. The idea of “Animal Spirits” means that these expectations are not ANCHORED by real variables – so essential parts of the causal mechanism driving the economy are in the minds of the people, and not easily quantified or captured by numbers. This discussion brings us to the end of our introductory textbook on statistics. In later, more advanced courses, we will discuss more about both the discovery of causal paths, and data analysis with ASSUMED causal paths.

For Subheading 2, use a subtitle such as “Why Today’s Kids Can’t Cope.” This subsection explains the differences between traditional and modern parenting styles, the role that technology has played in parenting, the decline in value systems, etc.

Chapter Summary/Key Takeaways

Insert content here...

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

- The TV and other digital devices today play the role of the parent in the home.
- Point 2 from your text...
- Point 3 from your text...
- etc.

In the next chapter, you will learn...

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter, link the next chapter's information with what has already been learned.

PART II: Insert Title of the Part

Use successive parts to cover the more detailed or complex areas of the book's topic. Since Part I defined the topic/problem, consider using this section to provide solutions. In this case, a suitable example title is "Identifying the Right Parenting Strategies" or "Bulletproofing Your Child's Mindset." Don't forget that the chapters in this part of the book must align with the Part title you have chosen.

Chapter Three: Insert Chapter Title

Begin a new chapter here...

For the purposes of this example, this chapter's title is "Developing a Bulletproof Mindset." This means that this chapter will be dealing with strategies for strengthening the mind. Start off by providing a brief overview of the information contained in the chapter and then transition smoothly into your supporting points. Try to keep the language simple and understandable to generate a rapport with the reader and keep them engaged.

Insert Subheading 1

Insert content here...

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are always related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is "Mental Discipline." Offer the reader practical strategies for training a child to develop fortitude, awareness, etc.

Insert Subheading 2

Insert content here...

For Subheading 2, use a subtitle such as "Self-Confidence." This example chapter provides practical steps and tips on how to train a child to become courageous, self-confident, etc.

Chapter Summary/Key Takeaways

Insert content here...

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

- Developing mental resilience is an important part of achieving success in life.
- Point 2 from your text...
- Point 3 from your text...
- etc.

In the next chapter, you will learn...

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter, link the next chapter's information with what has already been learned.

مَا أَصَابَ مِنْ مُصِيبَةٍ فِي الْأَرْضِ وَلَا فِي أَنفُسِكُمْ إِلَّا فِي كِتَابٍ أَنَّا نَبْرَأُهَا إِنَّ ذَلِكَ عَلَى اللَّهِ بِسِيرٍ ٢٢ لِكَيْلَ
٢٣ تَأْسُوا عَلَىٰ مَا فَانِتُمْ وَلَا تَفْرُخُو بِمَاٰءَاتِنَاكُمْ وَاللَّهُ لَا يُحِبُّ كُلَّ مُخْتَالٍ فَخُورٍ

No calamity 'or blessing' occurs on earth or in yourselves without being 'written' in a Record before We bring it into being. This is certainly easy for Allah. 'We let you know this' so that you neither grieve over what you have missed nor boast over what He has granted you. For Allah does not like whoever is arrogant, boastful

Chapter Four: Insert Chapter Title

Begin a new chapter here...

For the purposes of this example, this chapter's title is "Establishing Social Values." It covers the social/community aspect of a child's wellbeing. Start off by providing a brief overview of the information contained in the chapter and then transition smoothly into your supporting points. Try to keep the language simple and understandable to generate rapport with the reader and keep them engaged.

Insert Subheading 1

Insert content here...

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is "Getting Along with Others." Offer the reader practical strategies for teaching their child how to coexist with others, why it's important to do so, etc.

Insert Subheading 2

Insert content here...

For Subheading 2, use a subtitle such as "Developing Emotional Intelligence." It provides practical steps on how to help a child to read social cues, talk about any negative emotions, etc.

Chapter Summary/Key Takeaways

Insert content here...

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

- Instilling social values and ethics in a child will help them integrate well into society.
- Point 2 from your text...
- Point 3 from your text...

- etc.

In the next chapter, you will learn...

To logically transition smoothly from chapter to chapter, inform the reader of what is coming next. When ending your chapter with a paragraph, link the next chapter's information with what has already been learned.

Chapter Five: Insert Chapter Title

Begin a new chapter here...

Assuming this is your last chapter in the book, create a title that is somewhat forward-looking, for example, "Today's Kids, Future Parents." Here, present an outlook of how the strategies in this book, as well as new scientific insights, will shape the future. Also discuss what the world will look like if kids are not taught how to strengthen their mindset.

Insert Subheading 1

Insert content here...

Divide your chapter into sections with relevant subheadings. Subheadings guide the reader through the chapter and help in showing how you perceive the topic. Always have more than one subheading per chapter and make sure they are related to your chapter topic.

When researching content for a particular chapter, any key highlights you come across can act as a subheading. For example, Subheading 1 for this chapter is "Adapting to Change," where you then discuss the importance of developing adaptability and flexibility, etc.

Insert Subheading 2

Insert content here...

For Subheading 2, use a subtitle such as "Leaving a Legacy." Here, wrap everything up by discussing how all the strategies provided will ensure a better tomorrow for all.

Chapter Summary/Key Takeaways

Insert content here...

Remind the reader of the key points of the chapter in a short paragraph. Alternatively, use a bullet point format as shown below:

- In the future, survival will depend on the ability to adapt to a fast-changing environment.
- Point 2 from your text...
- Point 3 from your text...
- etc.

Epilogue/Conclusion

Insert content here...

Keep it short and sweet. Mention the key highlights of the book and the action steps to solve the problems. Remind the reader of how taking the said action steps will benefit them.

Bibliography

Below is an example of a list of works cited using APA style. Arrange your list of references alphabetically.

When citing books, use the format as shown in the following examples, applying Bibliography style to the format:

Author's last name, first initial or initials. (Publication date). *Book title*. Additional information. City, State of publication: Publishing company.

King, S. (2000). *On writing: A memoir of the craft*. New York, NY: Pocket Books.

When citing online resources, use the format as shown in the following examples:

For internet documents

Author's last name, first initial or initials. (Date of publication). Title of article. *Title of work*. Retrieved from full URL

Amir, N. (2018, October 17). 4 tips for staying on track with your writing. *Write Nonfiction now!* Retrieved from <http://writenonfictionnow.com/tips-staying-track-writing/>

For online periodicals

Author's last name, first initial or initials. (Date of publication). Title of article. *Title of Periodical*, volume and page numbers. Retrieved from full URL

Brewer, R. L. (2018, October 4). How to write better titles: 7 effective title tips for books, articles, and conference sessions. *Writer's Digest*. Retrieved from <http://www.writersdigest.com/whats-new/how-to-write-better-titles>

When citing magazines, use the format as shown in the following examples:

For magazines and periodicals

Author's last name, first initial or initials. (Publication date). Article title. *Title of periodical*, volume number (issue number if available), inclusive pages.

McPhee, J. (2013, April 29). Draft No. 4. *New Yorker*, 89, 20-25.

For more details and guidelines, consult the APA Publication Manual.

Acknowledgments

Insert content here...

Thank the key people who inspired you and helped you throughout the process of writing and publishing your work. This is somewhat similar to the dedication page, except here you can elaborate and include more people.

About the Author

Insert content here...

Write this page in the third person. (For example, use “the author” or your name, not “I.”) The information within establishes your credibility with readers. Avoid being too wordy. Simply provide your background and expertise on the topic of your book, as well as other information that will build trust. For example:

- Professional and personal achievements related to the topic at hand
- List of other published works and a link to your website
- Educational background
- Mention other notable experts in the field that you have worked with
- Your area of residence, family status, hobbies, etc.