

Lecture 1: Introduction

Before the study of economics, it is necessary to remind readers that never believe the ordinary textbooks of economics without thinking.

In general, textbooks are the truth. However, it is not the case for economics. There are many mistakes in the books of economics.

Why? Because economics is too young.

Let me introduce the history of Economics. A question: when was USA independent? The answer: in 1776. It is the same year that a British named Adam Smith (1723-1790) published his book "An Inquiry into the Nature and Causes of the Wealth of Nations," which marked the birth of economics. So Adam Smith is the founder of economics. Students and the researchers of economics, including you and me, are his disciples.

So you can see how young economics is, just like the USA that is only 200 years old. It is no wonder that there are lots of mistakes in economics.

Natural science, such as physics, is regarded as truth. However, don't forget that Aristotle started to study it thousands of years ago. It has been developing for thousands of years. There had also been many mistakes in physics. For example, physicists believed that the falling speed of an object was correlating with its mass, which is totally wrong. The falling rate of a stone is faster than the plume's just because the air friction interferes in. Another example is astronomy. How long had erroneous Geocentric been dominant in astronomy?

So it is not a surprise that there are many mistakes in economics because it is only 200 years old. It is a surprise that we entirely believe economics textbooks without thinking. In fact, many authors of economics textbooks do not take all that they write in books for granted. However, textbooks are not academic work, so it is better to follow the majority. Nevertheless, truth is often in the hand of the minority.

However, I decided to write a textbook on economics which does not follow the majority. I will select only the correct contents from ordinary textbooks of economics, criticize the mistakes and add something new and correct. Therefore, this book looks somewhat different from the usual textbooks of economics.

Someone would doubt whether the truth is subject to my judgment. How can we know the truth objectively? We need an assistant: Methodology of Science (MS) or Philosophy of Science which provides an objective analysis of whether a subject is a science. It applies not only for economics but also for all kinds of sciences. The scholars of MS establish the rules of science form the practice of natural science. So it makes one significant distinction between ordinary textbooks and this book that there are many introductions of MS in this book. In my opinion, the best way to master MS for those who are not the students of philosophy or logics is to learn economics.

Besides, this book is suitable for both students major in economics and amateurs. There is some math in this book for the economics students to study, while for beginners, math is not essential, though understanding it does help to understand economics better. But you do not need to be anxious if you cannot understand the math in this book, because the application of economics does not rely on math.

This textbook inherits the academic discipline from "Economics Explanation" by Steven N.G Cheung. The references of traditional economics textbook are mainly from "Theory of Price and Application" written by Cheung's teacher Jack Hirshleifer (1925-2005). Cheung wrote a preface "The Experience of Learning -- Preface for My Teacher's Textbook" for its Chinese version translated by Zhou Yan and me. In the preface, He suggested that students read the "Theory of Price and Application" with critical thinking instead of belief. It is more important to raise questions and discuss with classmates or teachers. That is the real study!

I would like to share these words with all the readers of this book as an encouragement!

Lecture 2: Economics is a Social Science

Economics is a social science, which is the subject nature of it. However, what does it mean? This lecture will explain it with the methodologies of science.

Section1: Economics is a Science – Methodology of Science (1): Refutability¹

Economics is a science. Science is an activity of human intelligence. It starts from one or several postulates or axioms from which a set of theories and inferences are established as a system. Besides, all the conclusions of this theoretical system must be tested by fact.

This definition of "Science" can be divided into three stages also known as "Trilogy of Science".

The first stage is postulates and axioms, which is the starting point of science. It does not need proof; neither needs to be true; it just needs to be accepted without any doubts. Sceptics doubt everything including those as a starting point. But what can be done without a starting point?

The second stage is a logical deduction. Mathematics as a tool assisting logical deduction may be useful for science. However, Math is not science itself, but a useful tool for the second stage of science.

The third stage is the fact test, which makes the distinction between science and math: there are only the previous two stages in math (Geometry is a typical instance). Without this third stage, math is not science. Since the first stage of science needs no proof and does not allow any doubts, it seems very arbitrary and even unreasonable. Then how can be it sure to be OK? That is why science needs the support of fact in the third stage.

It is the third stage that distinguishes science from math. Therefore, an essential feature of science is not provability, but refutability. That is, fact can refute it².

There is a serious and rather common misunderstanding about science that should be clarified. The public usually presumes that science is equivalent to truth. This is a misconception as they are totally different. A statement with refutability is scientific; otherwise, it is non-scientific. The false is the refuted but the truth is not the unrefuted. Therefore, it is not contradictory to say science is false and non-science is true.

Let us consider this example of unscientific truth: a four-footed animal has four feet. It is impossible to find a fact to refute this statement, which means it is absolutely true. However, what is irrefutable is exactly unscientific. Think about it deeply; why is that the statement cannot be falsified or refuted? Because the former part of it just repeats the latter part. In other words, A is B, so B is A. Such un-scientific, but true statement is called tautology which means repetition. Human study science in order to discover new knowledge, but tautology does not add any new knowledge; instead, it only repeats what already established. That is why tautology is absolutely true but becomes unscientific.

Tautology is not the only case of irrefutable statements. There are also four kinds of unscientific statements: unreality, ambiguity, contradiction and unrestrained.

The statement of "God exists." is unscientific; instead, it is associated more with religion. The question here is why so? This is because the concept of God is not a fact, and it cannot be tested by facts, which leads to irrefutability.

Seen from another point of view, the concept "God" is ambiguous, which also leads to irrefutability. Superstition, which is somewhat similar to religion, tries to avoid being refuted with ambiguity.

Here is a story: Three students go to take an important public examination. They meet a fortune-teller on the road, so ask him to predict their examination results. The fortune-teller just raises a finger and says nothing.

Possible outcomes: If one of them passes the examination, that one-finger means one succeeds. If two of them pass, it means one of them fails. If all of them pass or do not pass, it means they as a group succeed or fail³. Whatever the result will be, the fortune-teller can redefine the ambiguous meaning of the one-finger to make sure he is always right. However, it is unscientific precisely because it is irrefutable.

The statement of "There is a black point on an all-white wall." is also unscientific. With this contradiction, the conclusion of "The wall is a god" will be deducted. Is it only an omnipotent god who can be both black and white?

In a math class at middle schools, they teach that a statement and its contrapositive are logically equivalent. However, only a few know that it is also the way of testing science.

If " $A \Rightarrow B$ " is tested, then " $\text{not}(B) \Rightarrow \text{not}(A)$ " instead of " $\text{not}(A) \Rightarrow \text{not}(B)$ " should be used.

By analogy if " $A \Rightarrow B$ and C " is tested, then " $\text{not}(B) \Rightarrow \text{not}(A)$ " and " $\text{not}(C) \Rightarrow \text{not}(A)$ " should be used. If " $A \Rightarrow B, C, D\dots$ " is tested, because there are limitless propositions deduced from A, the test about A are endless, and become irrefutable. It is not easy to find a simple case to illustrate this situation of unrestrained, which will be pointed out when it occurs.

"Scientific faults" are refutable statements and have been refuted by facts. For example; there are two conflicting theories of Heliocentric and Geocentric on the relationship between sun and earth.

Which of them is true? How to test them? Copernicus argued that if Heliocentric were true, then Venus would be waxing and waning like the moon, which is " $A \Rightarrow B$ " where A is Heliocentric, and B is the Venus' phase. So the test condition for this statement is that if the Venus were not waxing and waning, Heliocentric would be false, which is " $\text{not}(A) \Rightarrow \text{not}(B)$ ". Heliocentric can present a refutable condition, so it is scientific. On the other hand, Geocentric can also present its refutable condition: if the Venus were waxing and waning, Geocentric would be false, so it is scientific too. Afterwards, Galileo observed with a telescope, found that the Venus does be waxing and waning, which refuted Geocentric and supported Heliocentric. Then Geocentric is scientifically false because it has been reputed though it is reputable. By contrast, Heliocentric is not only refutable but also true, because the observation of Galileo does not refute it. "Reputability" is the characteristic of science, while "Refuted" is the end of false theories, which should not be confused with each other.

Thus human observes the reality, exact the laws out of facts, then make up science theories which can be used to explain what has happened and predict what will happen. If the explanations and predictions have not been refuted by facts, then theories can be accepted as true. But if they are refuted, the theories need to be revised, or even abandoned entirely.

The example of Heliocentric illustrates that the theories accepted as true are not true. More accurately speaking, they are just not refuted for now, which does not mean that they will never be refuted in the future, because they are always refutable in science. In fact, with the development of Astronomy, Heliocentric theory has been revised from the idea that all planets, including the Earth, move around the Sun from the centric orbits to the oval orbits. Now the heliocentric theory has been abandoned and replaced by the theory of cosmic expansion because the sun is no longer believed as the center of the Universe.

Logics also deny that science or refutable theories can be absolutely true. For example, even human has seen thousands of white swans and find no one is black; still, they cannot conclude that there are no black ones in the world. The statement "all swans are white" can be accepted as true, but just for now.

Of course, only scientifically true theories are desirable, because they can increase human's knowledge and guide the practice. As mentioned above, it is the same for explaining what has happened and predicting what will happen. In practice, human firstly tries to explain what has happened to construct theories, then with which predict what will happen by observing if the predictions come true, human test the theories again and again.

It seems rather abstract, so let us take examples to illustrate the above interpretation. In Physics, Newton's laws explain or deduct the phenomenon of an apple falling on the ground. They are accepted as true because we never see an apple flying to the sky. Airplanes that can fly to the sky are invented, which does not violate but apply Newton's laws. As long as the conditions required by Newton's laws are satisfied, airplanes never fall from the sky, which proves the laws are true.

So, why is Economics scientific?

Economics is a set of systematic theories deduced from a postulate which is "human are all self-interested", and an axiom which is the Law of Demand. Economics can be used to explain some phenomena that cannot be explained by Physics. For example, there is a \$100 bill in a public place. Physics can explain why it is gone with the wind. However, it cannot explain why the bill is still gone without wind which only can be explained by economics.

The Theory of Supply and Demand (TSD) introduced in later lectures argues that the price will rise if the demand exceeds the supply, while the price will fall if the supply exceeds the demand. As long as the conditions required by the TSD are satisfied, it has never been refuted. If the government imposes price control, forbidding sellers to increase prices when the demand exceeds supply, it seems that the price does not rise, but it does in the other form, such as the time spent on queuing and the red pockets paid as bribe. The logic is similar to the fact of airplanes flying to the sky does not violate Newton's laws. So, it just proves that the TSD is true again.

1. This section introduces the key points of MS. Those who are interested in it should read such masterpieces as "The Logic of Scientific Discovery" (Karl Popper, 1902-1994) and "The Philosophy of Natural Science" (Carl G. Hempel, 1905-1997). [↗](#)

2. Here "Science" is a narrow-sense concept. Some scholars call those subjects without refutability, such as Logics and Mathematics, as "Formal Science", while those with refutability as "Empirical Science". That concept of science is a broad-sense one. This book use "Science" in narrow-sense which is refutable, i.e. "Empirical Science". [↗](#)

3. In Chinese, the word "one" can be used to mean a person or a group of people altogether. [↗](#)

Section 2: Economics is a Social Science

Economics is a science, but it is a social science different from natural science such as Physics.

What is the difference? Physics studies matter, while economics studies human, which is the fundamental difference.

Why is it a fundamental difference? Because economists are human too! In other words, economists themselves are also the study objects. Chemistry is a natural science, of which the study object is a molecule, while the studiers are human. What happens if molecules study themselves? Since economists as studiers and their study objects are all human, it causes much trouble.

Firstly, human has emotion, while science requires objective analysis where subjective emotion should be ruled out. It is easy for natural scientists with the matter as a study object to do so, while rather tricky for social scientists who themselves are also human to rule out their emotions. Economic issues often arouse social disputes with anger. Anger is precisely a kind of emotion which should be ruled out in the discussion of economic issues since economics is a science. Anger can never be used to discuss issues of physics; however, once issues are economical, it is difficult to rule out the emotion of anger!

Secondly, the postulate of economics is the self-interest of human. Economists are, of course, self-interested too, which cause a severe problem: some economists do not objectively analyze issues with economic theories, but distort their analysis intentionally or unintentionally and make the conclusions beneficial to them. For example, there are many debates on the issue of property prices in China. Different economists have very different opinions. No wonder the public doubt if economics is a science.

There is a joke that ten economists would have eleven opinions on the same economic issue.

There are three reasons for this weird phenomenon:

(1) As mentioned in the previous lecture, there are still so many mistakes in economics nowadays. Only wrong conclusions can be made by analysis with wrong theories.

(2) Economics is rather simple but very difficult to apply. So there is a great distinction between a master and a lame duck. A skillful woodcarver can make a masterpiece out of wood while a clumsy one cannot make anything. Similarly, the second-rate or unqualified economists can only make the wrong conclusion even with correct economic theories.

(3) It may be the most key reason. As mentioned above, the self-interested economists distort their analysis to make the conclusions most beneficial to them. It is neither because the theories are wrong, nor because they do not understand how to use the theories correctly, but because they need these conclusions beneficial to them.

Natural science such as physics has the advantage that it can avoid the distortions by self-interest and emotion of human. However, this kind of distortion did occur in the history of the developments of natural science. It may be not well-known that Newton was not a good guy. Historians believe that the fight for the invention of calculus between Newton and Leibniz had postponed the development of mathematics in the UK by at least a century. Besides, he was so narrow-minded that he severely squeezed Hooker, another famous physicist, out.

There is a real story in "The Passions of the Mind", a biography of Freud. When Freud began his career as a psychiatrist, in medicine, there was a wrong thought and people believed that only females would suffer from hysteria because they thought psychosis was only for those with weak will. Due to this discrimination for female, psychiatrists at that time, never thought about the possibility of male patients.

However, Freud found some cases of male patients of hysteria in the clinic. Napoleon was a patient of hysteria, which showed that hysteria had nothing to do with a weak will and that even males can get hysteria. Nevertheless, at that time, the prejudice for female was so strong that such ridiculous thought was prevalent even though medicine is natural science with the human body instead of a human being as its study object.

Freud published his finding in an academic paper which caused criticisms and attack on him overwhelmingly. He did not take them seriously but was surprised and upset by the criticism of a massively respected elder in medicine whom he adored. He felt puzzled why the elder who is more experienced than him refused to admit that there were male patients of hysteria.

Due to the intense criticisms, Freud lost his reputation and even job. Many years later, when the respected elder was dying, he called Freud to his bed and said: "In fact, no one in the world knows more clearly than I that you are right, because I am a male patient of hysteria myself."

The respected elder was not ignorant. Precisely because he knew undoubtedly, he could not accept the truth. If one refuses to accept the truth because of ignorance, he cannot be educated. Nevertheless, if the truth is adverse to his benefits, it is in vain to try to persuade him by reasoning. There is a Chinese saying that a scholar cannot reason with a soldier. That is because the soldier does not intend to reason with him. It is the same for those who want to hide their selfish desires by disguising to reason.

Thus natural science with the matter as study object has the same problem as economics. However, natural science has the advantage that it can do controllable experiments in labs to test theories. Different persons have different interests, so it is usually not easy for natural science to be manipulated by a few persons. While the lab of economics is the uncontrollable real world, so it is much easier for few economists to take advantage of the public's ignorance and the emotion to distort the economic analysis and make conclusions beneficial to them at the expense of the reputation of all economists as a whole.

The public is confused to see different economists have such divergent opinions on the same issue hence doubt if economics is a science like physics. It is some economists instead of economics that should carry the blames.

READING MATERIAL: THE ECOLOGIST, THE ECONOMIST, AND THE STATISTICIAN¹

In 1990 the ecologist and popular author Paul Ehrlich sent a check for \$576.07 to the economist Julian L. Simon. It was a payoff on a bet made 10 years previously. Ehrlich, though not an economist, had made startling economic predictions in his 1968 best-seller *The Population Bomb*. Take, for example, his opening sentences: "The battle to feed all of humanity is over. In the 1970s, hundreds of millions of people are going to starve to death."

Ehrlich's predictions utterly failed. Despite an "oil crisis" in mid-decade and again toward the end, overall the 1970s were a decade of remarkable economic growth. Yet the author continued to be highly acclaimed. In books, speeches, and articles that received worldwide attention, he prophesied that accessible supplies of many key minerals would be nearly depleted by 1985.

Meanwhile the economist Julian L. Simon had been predicting just the reverse for the 1970s and 1980s: "continuing improvements in human well-being, and lower prices for raw

materials. Simon challenged Ehrlich to back his contrary forecast with hard cash, and Ehrlich accepted the challenge. The result was a 1980 bet about whether the prices of five important metals – chrome, copper, nickel, tin, and tungsten – would, after allowing for inflation, rise or fall by 1990. The economist let the ecologist choose specific commodities". Once again, economist Simon was proved right and Ehrlich wrong. The 1980s were also a decade of prosperity, and raw materials prices generally fell. The ecologist had to pay off on the bet.

Ehrlich's analytic error was to look only at the demand side, and specifically at a growing world population (more mouths to feed). Julian Simon, with his better economic understanding, also considered the supply side. More people mean more mouths, but also more hands and brains. Simon also took account of other favorable economic trends such as greater liberalization of national economies and increased international trade.

So the predictions of economic analysis were vindicated. Nevertheless, Paul Ehrlich continued to publish highly popular books forecasting, as mistakenly as before, environmental doom in the near future. Just as regularly, up to his death in 1998 Julian Simon was providing sound analysis and well-validated economic forecasts, culminating in his volume *The Ultimate Resource* (2nd ed., Princeton University Press, 1998). But none of his books was best-sellers.

What was going on here? An economic interpretation might be that these two authors were supplying different commodities. Julian Simon was supplying correct economic analysis. Paul Ehrlich was and is in a different business, somewhat comparable to horror-story writers such as Stephen King. Evidently, at least when it comes to selling books, the demand for horror stories far exceeds the demand for sound economic analysis. [For a detailed history of the Ehrlich-Simon wager see John Tierney, "Betting the Planet," *New York Times Magazine*, December 2, 1990.]

A later exhaustive and careful study by the Danish statistician Bjorn Lomborg (*The Skeptical Environmentalist*, Cambridge University Press, 2001) has updated Simon's results. Initially intending to refute Simon's analysis, Lomborg found his research instead of confirming it. Despite the rising world population, per capita incomes have maintained their upward trend and real prices of essential resources have continued to fall.

As a rather strange sidelight, upon publishing these results, Lomborg became the target of personal attacks that reached an extraordinary level of intensity. At one of his public presentations, a cream pie was dumped on his face and clothing. He was even accused of "intellectual dishonesty" by a scientific body with official standing in Denmark. Ultimately, all the charges were conclusively refuted, and the improperly motivated proceeding against Lomborg was dismissed. (Which does not necessarily mean, of course, that all of Lomborg's estimates and forecasts will ultimately prove to be correct.)

One possible lesson: sound economic analysis can sometimes be riskier than one might think! ²

The material above shows that Ehrlich was not an economist, and facts had refuted his analysis, but his books were still more popular than those by Simon as an economist. Why? Because the demand for Ehrlich's book is the same as for horror fictions instead of correct economic analysis. Ehrlich kept writing books full of mistakes, not to provide correct economic analysis, but to attract readers as an alarmist. There are similar cases in China too. Who should carry the blames, economics, or some so-called economists?

However, it is necessary to point out that when people enjoy themselves by reading horror fictions, they know that they are just fictions, it will not feel panic. Who would believe that the film of "2012" would become real, and be so horrified to sell out their properties or even commit suicide? But when some economists, with the name of a scientist, are engaged in faking horror fictions, the consequences will be severe. The public thought what they provide were correct economic analysis, but they only make alarmist remarks to earn money. If the government does believe them and implement economic policies according to the analysis of such economists, the consequence will be disastrous.

1. This reading material is an excerpt from "Price Theory and Its Application", the seventh edition, Jack Hirshleifer, Amihai Glazer, David Hirshleifer, Cambridge University Press, 2005, P6-7. This annotation applies to the following materials which are also from this book, so will not be repeated. [↗](#)

2. A summary of the Lomborg controversy is provided in Jim Giles, "The Man They Love to Hate," Nature, v. 423, May 15, 2003. [↗](#)

Section 3: Why Do We Need to Study Economics?

Why do we need to study Economics? Many readers may believe that those individuals major in economics are easier to find good jobs (and jobs with high incomes), or learn to know how to earn money. In reality, those whose careers are economic-related do have higher incomes, but they are not the richest ones while the richest ones may not be those major in economics or even have never learned economics.

As an economics teacher, I hope that the readers would master the way of thinking like an economist as most are not, and will not be economists. They do not need to be economists, but they need to think like economists.

As mentioned above, economics is a science, analyzing problems objectively just like natural science. However, natural science only teaches us how to analyze problems about the matter, while economics teaches us how to analyze the problems about human. To learn to think like an economist is to learn to analyze problems objectively, ruling out emotions like anger, as this should be economics most benefit to ordinary people.

Besides, some economists distort economics analysis for their interests and try to mislead the public and governments to benefit themselves. Therefore, if one has learned economics, he will not be fooled by such economists. Mrs. Robinson, a famous British female economist, once said: "I study economics in order to avoid being deceived by economists." That is also my Motto.

Lecture 3: Self-interest: the Postulate of Economics

In the last lecture, we have clarified that economics is a science. Science starts from one or some postulates and axioms, and through logical deduction constructs a theoretical system that can explain facts. Therefore, in this chapter, we are going to introduce the postulate of economics.

Section 1: Self-interest

The postulate of economics is "Self-interest" or more grandly "Rationality".

The postulate of Self-interest means that human will try to achieve the best result with the lowest cost. It can be regarded as a biological instinct of animals. Under different circumstances, "the least cost" can be equivalent to the lowest price of money spent or the minimal resources consumed, while "best results" can be equivalent to the maximal income or revenue.

Caution: self-interest does not mean to benefit oneself by harming others unless harming others is the least costly way to realize one's self-interest goal. In fact, in most cases, in order to benefit oneself with the least cost, one can better act altruistically. As mentioned in Lecture 1, there is a famous example in the book, *Wealth of Nation* written by the founder of economics, Adam Smith. It says: the baker bakes bread for his good. He wants to earn money from buyers, but as a result, it also meets their needs. So, subjectively he is doing the job for his good, but objectively he is providing us with food, which benefits himself and others at the same time. He could rob people to get money, but that would put himself in jail. Also, he could beg for money, but few people would be willing to give beggar money for nothing, which means begging would not earn him as much as selling bread. After comparing the costs of the different ways of getting money, he chooses the one with the lowest cost and the highest income, which generally is selling bread.

Of course, if the place where baker lives had poor public security, robbery would become easier, because the risk of being caught would not be high enough. When committing crime costs less than selling bread, the baker would change his career path. Thus, the choice between obeying the law and committing a crime is also the result of selfish-interest by measuring incomes and costs. Actually, beggars are said to have a higher income than bakers; however, not too much people are willing to be beggars, because bakers are more respectable, which means that the so-called income should not narrowly be considered as monetary income, but also contains reputation, dignity, and other non-monetary income. So, incomes include monetary ones and non-monetary. The same applies to costs. If one does not choose to be a beggar, the total incomes of beggar must be not higher than those of baker, and the total costs are not lower, which implies that self-interest dominates his behavior.

Let us think about this phenomenon more profound. Why do bakers earn more non-monetary incomes than beggars? Due to morality, bakers are respected while beggars are belittled. However, why is morality as such? Because bakers contribute to others and create new wealth by providing bread, while beggars do not contribute to others and create no wealth for society. They transfer wealth from those who donate. In other words, morality guides human to realize self-interest by altruism. That is precisely the reason why human need morality. Morality rewards self-interest acts by altruism, belittles those without benefits to others (such as begging) and condemns those by harming others (such as robbing). Therefore, robbery is not only sanctioned by law but also bears the pressure of moral condemnation which is a non-monetary cost. So morality increases the total costs of robbery. This is the same as the market guiding human to behave altruistically by price and the law preventing them from harming others by penalty.

Traditional morality makes it shameful to be a beggar, but modern social welfare essentially encourages people to be beggars. The original intention of social welfare is to help the poor, but the real effect is to support the lazy. The reason is straightforward: for one who is not sick or disabled, without the belittling of morality as a non-monetary cost, why not just sit there and live on relief. The national debt crisis of Greece tells this truth: the Greeks are not born to be lazy, and it is the long-term social welfare that makes them so. Also, it corrupts morality which praises self-

reliance and belittles idleness. People feel no shame in taking it for granted to ask for money from the government, and scold government no matter it refuses or gives less than before.

Section 2: The Postulate Needs not to be True

Someone may argue, "I am not self-interested. It is wrong that economics postulates that one is self-interested." However, the postulates and axioms of science do not need any proof, which means they do not need to be real or genuine. However, they do need to be tested by facts. It means the theories deduced from them can explain phenomena in reality. If they can't, the postulates and axioms must be modified or even rejected. Therefore, what can reject the postulate of self-interest is the explanation power of theories deducted from it instead of the truth or correctness of it.

Some experimental economists try to prove that the postulate of self-interest or rationality is wrong. However, they make an elementary logical mistake. In Lecture 2 about the methodology of science, it has been pointed out that if " $A \Rightarrow B$ " is tested, then " $\text{not } (B) \Rightarrow \text{not } (A)$ " should be used. One cannot conclude that " $A \Rightarrow B$ " is wrong because " $\text{not } (A) \Rightarrow \text{not } (B)$ ". Thus it is also seriously wrong to refute economics by experiments which refute the postulate of self-interest or rationality.

Let us consider this imaginary example to show that even if the postulate of self-interest or rationality is wrong, it can still explain or predict correctly.¹ There are a group of people who are going to build gas stations, and there are three places available: the highway, the mountain and the desert. Suppose the construction costs are the same in these three places. According to the postulate of self-interest or rationality, economics predicts that they will build gas stations along the highway. Actually they are idiots, so it is totally wrong to postulate them to be self-interested or rational. Nevertheless, it does not matter! These idiots build gas stations just at random, so some of them build the gas stations in the mountain, while some of them in the desert. Of course, some idiots happen to build the gas stations along the highway. After some time, gas stations in the mountain and dessert go bankrupt because income earned cannot cover operational costs while those along the highway are prosperous and survive more in the business. The final result is the same as economic prediction. That is enough! It is not important whether they choose the place to build the gas stations by careful calculation or at random. It is essential that finally there is no gas station in the mountain or the desert, but there are along the highway, so facts have not refuted the economic prediction.

Milton Friedman (1912-2006) gave another similar example in his book "Price Theory". Why do plants grow toward the sun? Biology has a theory to explain this phenomenon. The growth of leaves needs sunlight for photosynthesis, so more leaves are growing in the direction toward the sun and vice versa. As time passes, plants appear to grow toward the sun. Economics can also explain this phenomenon. How? Plants are postulated to be self-interested or rational, and the sunlight is beneficial to their growth, so they grow towards the sun. It is wrong to postulate plants to be self-interested or rational. How can plants think? It does not matter. The conclusion of economic explanation can be the same as that of biological explanation.

There is a new inter-discipline called bio-economics which uses the postulate of economics to deduce the theories explaining the biological phenomena. Moreover, the biological study also supports the postulate of economics. There is a famous book "The Extended Selfish Gene" in biology. It is well-known that species evolve through genetic mutations. Some mutations conform to the nature of selfishness, while some do not. Those genes without the nature of selfishness will be eliminated in natural selection, while those who are selfish survive. The survivors keep duplicating their genes and multiply, which makes selfishness continuously strengthened and

inherited, thus ultimately determines that all species, including human, are selfish. Isn't this process similar to the example of building gas stations by the idiots? Genetic mutations are random, just like the choices of the idiots. Nevertheless, natural selection determines that only the gas stations or genes conforming to the nature of selfishness can survive. It is well-known that Darwin's idea is natural selection or survival of the fittest, but few know that the exposition inspired him about the self-interest of human in Adam Smith's "Wealth of Nations".

However, it is not crucial whether a human is self-interested or not, or whether he is selfish due to the genes as "The Extended Selfish Gene" suggests, because self-interest is a postulate which does not need to be proved but must be accepted. If one does not accept it, then it is better not to study economics. Of course, one can invent a subject with the postulate of un-self-interest. Those who study this subject must accept it, but the theories deducted from it need to be tested by facts. If facts do not refute the theories, one will succeed in establishing a brand-new subject called Economics of Un-self-interest, while the economics now may be renamed as Economics of Self-interest. However, if facts refute one's theories, he has to modify or even abandon the postulate of un-self-interest. No one has been able to do so, because the theories deducted from the postulate of un-self-interest are easily refuted by facts and cannot be established as a subject. In geometry, however, it did happen that new subjects were established by changing postulates. The geometry learned in middle school is "Euclidean geometry" because ancient Greek Euclid established it. There are five postulates in Euclidean geometry. Among them, the axiom derived from the fifth postulate is called "the parallel axiom" that "in a plane, through any point not on a given line, only one new line can be drawn that is parallel to the original one". However, in the 19th century, a Russian adopted a contradictory axiom that "in a plane, through any point not on a given line, at least two new lines can be drawn that is parallel to the original one." As a result, a new set of geometry theories without logical contradiction within it was deduced based on this new axiom. Later on, another German also deduced another set of new geometry theories without logical contradiction within it based on another contradictory axiom that "in a plane, through any point not on a given line, no new line can be drawn that is parallel to the original one." Eventually, the geometry based on the parallel axiom is called Euclidean geometry, while those based on alternative axioms other than the parallel one are collectively called non-Euclidean geometries.

Science is like this: In the first stage, there are one or several postulates which need not be proved, nor need it be true or real, but all in this discipline must accept and abide by the postulates. Anyone can establish new subjects based on different postulates, but whether they are successfully established must be tested by facts.

1. This imaginary example is adapted from "The Science of Demand", the 1st volume of the 1st edition of "Economic Explanation" by Steven N. S. Cheung, The Arcadia Press, May 2001, P44. [↩](#)

Section 3: The Postulate of Self-interest Must be Followed

Since economics has postulated human to be self-interested, it can not postulate human to be un-self-interested at the same time. In other words, one can not explain human behaviours with un-self-interest, or there can not be cases of un-self-interest. It is the same as Geometry: there can not be any theories that are contradictory to the parallel axiom in Euclidean Geometry. It is also the same as laws. The constitution is the fundamental law of a nation just like postulates and axioms of science. There can not be any laws that are contradictory to the constitution, or else they must be amended or even abolished.

What would happen if economics postulated human to be both self-interested and un-self-interested at the same time? If it was the case, then economics could be able to explain any phenomena, but actually, economics had not been able to explain it all.

For instance, why does one cheat others for money: It is because he is self-interested. Moreover, why does one donate money to others? Because he is un-self-interested.

Economics seems to have explained both the behaviors, but neither has been explained. When would on earth one be self-interested, and when would he be un-self-interested? How to predict if one would cheat or donate money in advance? As mentioned before, it is the same for explaining what has happened and predicting what will happen. In practice, theories are more needed to help to predict the future. If human were postulated to be both self-interested and un-self-interested, it would be impossible to predict their behaviors and be impossible to explain when they will cheat or donate money. This scenario brings out the specific case of tautology.

However, it would also be false to answer like that: why does one cheat others of money? Is it because he is self-interested, or why does one donate money to others? Is it because he is un-self-interested?

Self-interest is a postulate which can not apply directly to explain phenomena; otherwise, the explanations with it are directly postulated. It would happen again that any phenomena could be explained, but actually, they had not been explained at all, which is also the case of tautology. This scenario shows how it is easy to make the mistakes of tautology!

It is relatively easy to explain why one chooses to cheat. As mentioned before, when committing a crime was costing less than selling bread, the baker would change his career path. The explanations for both phenomena are similar. If cheating is an easy way of getting money, and would not be caught by the police, one would tend to do so.

How about donating? If the donee is a rich kid and he can go to school even without donations, one would not tend to donate to him. In contrast, if the donee is a poor kid and he cannot go to school without donations, one would tend to donate to him.

Furthermore, if the poor kid has no intention of study and donations would not make him study well, one would not tend to donate to him. In contrast, if the kid has the passion and intelligence to study and just lack of money, one would tend to donate to him.

Still, furthermore, suppose there is a charity and one needs to go a few miles to donate, one would tend to do so. In contrast, if one has to drive or take a car to the backwaters thousands of miles away to donate directly, he would not take such trouble. If there is a severe charity scandal exposed, one finds out that his donations have been abused instead of helping the poor, he would not tend to donate again.¹

The effect of a charity scandal on donations can be used to demonstrate the methodology of science introduced in the previous lecture. Suppose there are three kinds of economics: Economics of Self-interest, Economics of Un-self-interest, and Economics of Sometimes Self-interest and Sometimes Un-self-interest. A charity scandal will be used to test which of them is scientifically valid.

According to the postulate of Economics of Self-interest, if there are scandals, then donations to charity will decline, but if they increase, then Economics of Self-interest is refuted. This situation can present a refutable condition, so it is scientific. Donations do drop dramatically after the exposure of the charity scandal, so it is scientifically true.

In contrast, according to the postulate of Economics of Un-self-interest, donations to the charity will increase due to scandals, but if they decline, Economics of Un-self-interest is refuted. It can present the refutable condition, so it is scientific. However, the donations drop after the exposure of the charity scandal, so it is a scientific fault. It is too easy to refute Economics of Un-self-interest by facts; no wonder no one has ever tried to establish such economics.

How about Economics of Sometimes Self-interest and Sometimes Un-self-interest? Can it represent the refutable condition? No way! If the donations to the charity decline, it will say because human is self-interested now; while if they increase, it will say because human is un-self-interested now. It is always true, no matter what happens. However, precisely due to its irrefutability, it is a tautology instead of a science. One may doubt if there would be such ridiculous economics. Of course, there is, but with higher-sounding names. Earlier there was a concept of "bounded rationality" that prevailed for a while, and the idea of "irrationality" has been alive until now. Besides, there is "behavioral economics" which is rather popular and hot at present. The nature of all above is Economics of Sometimes Self-interest and Sometimes Un-self-interest. There are full of such non-scientific concepts and ideas in mainstream economics; no wonder it is despised by natural scientists.

In short, although donating is a kind of philanthropy which seems to be contrary to the postulate of self-interest, in fact, it is dominated by self-interest. Self-interest means that people will try to achieve the best results with the least cost. When one wants to help others, he still wants to achieve the best result for this goal with the least cost. So it is better to donate to poor kids rather than to rich ones. It is better to donate to the hard-working kids other than to those without the intention of studying. Moreover, it costs less to distribute the donations by a reputable charity. However, if the charity is not reliable at all, it is better to send directly to the donees by oneself. Thus, the fact that human is altruistic has not refuted the postulate of self-interest. Because even if altruism is the result one wants to achieve, he still wants to achieve it with the least cost, which means that self-interest still dominates his behaviors.

The postulate of self-interest can not directly apply to explain donating, but we can show how self-interest dominates the choices of a donator under different constraints. "Constraint" is a fundamental concept in economics, meaning the objective restriction on human's behaviors. Self-interest is also a restriction on human behaviors, but it is a postulate, needs not to be true and not an objective fact, while constraint must be.

The postulate of self-interest can not directly apply to explain human behaviors in economics. Instead, we need to point out different constraints and predict what a self-interested person will do under such constraints. This is both a prediction and an explanation. Finally, we need to investigate whether the person does do so as we predict. If he does, that means facts have tested the postulate of self-interest; otherwise, it has been refuted.

Neither can the postulate of self-interest directly apply to explain robbery or cheating. Different constraints are: one is that the police are powerful, and robbers and cheaters will be caught and punished. Under such a constraint, self-interested people will not choose to commit crimes. The other is the opposite situation, and it is more profitable to commit a crime.

In conclusion, there are different choices or behaviors of human not because human are sometimes self-interested and sometimes not, but because of different constraints. We can not apply economics to explain what has happened or predict what will happen by changing the postulate of self-interest, but investigating the change of constraints.

1.

Some may argue that there is a trick with the concept of donation in the explanation. Upon donating, the question is why one chooses to donate? Rather than why one chooses to donate to a poor and hard-working kid after he had chosen to donate.

Indeed, there are two aspects of questions about donation. Firstly, why does one choose to donate, or why does one not keep the money for his use? Secondly, having chosen to donate, how does one choose to donate with the least costs and best result. The explanation here is the answer to the second question. The first question has been answered when it is discussed why human needs morality in the first section of this lecture.

More specifically, suppose there are two persons in a society, and one is poor while the other is rich. If there were no morality, the rich one would only think about himself, while the poor one would die of starvation. Then only the rich will survive, but the question is how long he could survive in competition with other animals after he is alone in the society? However, if there is morality, the rich would donate his surplus to the poor one who would not die of starvation, and thus the competition with other animals will be minimal, and a chance for survival will be higher in comparison to if he was left alone.

This example shows that morality is precisely the result of selfish-interest of all human beings as a whole who tries to maximize the chances of survival in competitions with other species.

«

Section 4: Rationality is not Equivalent to Correctness

In Mainstream Economics or textbooks, there are often mistakes of being a conflict with the postulate of self-interest, because there are so many theories about "irrationality". For instance, it is said that there is irrational exuberance when the stock market is rising too much. As mentioned at the beginning of this lecture, the meaning of self-interest and rationality is the same. If one is said to be irrational, it means he is un-self-interested, which is a conflict with the postulate of economics, so it is a serious mistake. Some economists would be cautious about saying that it is difficult to judge how much the stock index could be regarded as irrationality. However, it is not the problem of how to judge if the stock index is irrational or not. It is forbidden in economics to use irrationality to explain human behaviors, which is the same as it is forbidden in non-Euclidean geometry to use the parallel axiom.

No matter how crazily the price of a stock rises, it is rational. That sounds too ridiculous to accept. Nevertheless, the critical point is: that human is self-interested or rational does not mean they are sure to make correct choices. The crazily rising of a stock price is the result of the investors' rational choices which may not be correct. Are rational or self-interested choices correct? Regardless of the cases mentioned before that committing crimes is rational but morally wrong behavior, here is another case. One is rational to make a choice, which means he believes he can achieve the best result with the lowest cost, but he may find he is wrong in the end. In other words, he has not achieved the best result or paid a higher price.

Take the stock investment as an example. One must believe the price of the stock he is going to buy would rise. However, the price of the stock may fall after he bought, so he suffered a loss. Was he irrational? No, he only made a mistake in his expectation of the stock price.

Moreover, why did he make a mistake? Because he is not the omniscient God! He thought that the stock price would rise, but it falls eventually. If he could know the price would fall in advance, of course, he would not choose to buy the stock. It is not because he is irrational, but because he does not know! If he knew the price would fall in advance, but he still chose to buy the stock, then the postulate of rationality would have been refuted by facts. That has never happened.

Human is rational, but he is not omniscient, which means ignorance is a constraint. Some have reliable inside information of the stock market, and know more than others or are less constrained, so their choices are more correct than those who are relatively ignorant. They are not more rational than others, but just under a different constraint. It is always the difference of constraints rather than rationality or irrationality that causes different human behaviors and their results.

Let us consider another example in reality for illustration. A student of economics posted in an economic forum that a classmate of his had not been going to take the postgraduate entrance examination. However, later he heard that someone not as good as he had passed the examination, so he completely changed his decision. The poster felt doubt that his classmate changed his decision just due to hearing such a piece of news. Is it right for economics to postulate human to be rational?

Has it refuted the postulate of rationality? Of course not. The classmate was not going to take the postgraduate entrance examination, of which he must have estimated the cost and benefit. He had to choose to spend time preparing for the exam or job hunting. Once he failed in the exam, neither had a job, he would suffer a significant loss. In the beginning, he estimated the exam to be

very difficult. It must be because of the information he had that made him believe that he had little chance to pass. He rationally gave up the exam after the calculation the cost and benefit. Later, he got some new information that someone not as good as he had passed the exam, so he adjusted his estimation about the exam, which means the constraints have changed. Thanks to the new information, he now knows better about the difficulty of the exam or becomes less ignorant than he was before. Along with the change of constraints, his choice or behavior also changed. If he knew that a postgraduate degree was better than a graduate job and he could pass the exam, but he still did not change his choice, then we would have arrived in conclusion that postulate of rationality has been refuted. Remember: it is always the change of constraints other than the postulate of self-interest or rationality that can be used to explain human's behaviors.

Some economists suggest that the reasons why self-interested or rational people cannot make correct choices are divided into ignorance and uncertainty. However, they are the same because uncertainty means it is impossible to completely know the future, which is equivalent to the ignorance of the future.

Ignorance is also a constraint which is defined as transaction cost in economics. Transaction cost is a fundamental but complicated concept, so it will not be explained in detail until the later lecture in this book. However, this concept is so indispensable that we need to frequently apply it to almost every issue of economics before that lecture. A reader can take his time to comprehend it by observing how it is applied.

One kind of transaction costs is information cost which is the cost of acquiring information. Ignorance is precisely due to information cost. One would choose to be ignorant rationally when there is too high information cost. Ignorance is a rational choice! One can benefit from knowing more, for which he has to pay formation cost. If the information cost is higher than the benefit, it is not worth knowing more, so it is rational to choose to be ignorant. When one becomes more enlightened, he is constrained less by information cost.

As mentioned above, uncertainty is the ignorance of the future. Human has made great efforts to predict the future, including weather prediction, earthquake prediction and even such unscientific means as fortune-telling. Great efforts imply tremendous information costs. To sum up, information cost is one kind of transaction costs, and transaction cost is a significant constraint that would affect human's behaviors.

Section 5: The Consequences of Ignoring the Postulate of Self-interest

Economics postulates human to be self-interested. It does not matter whether it is true or false. What does matter is facts have not rejected the results deduced from it. If the government does not take this postulate into account, or even just not carefully enough, there would be unexpected effects of policy implementation which would even run counter to the original intentions. For example, social welfare is intended to help the poor and the weak, but it makes them lazy and even corrupts morality. There are such counterproductive policies everywhere in reality. The policy-makers should be cautious: do not just think that the original intentions are good, but try to have a guess what the self-interested people would respond with a countermeasure.

READING MATERIAL: RATIONAL DRIVERS? 1

Air-bags reduce, on average, the severity of auto injuries and the risk of death from motor accidents. However, given that additional safety margin, it might well be rational for motorists to drive more aggressively!

Is such an adaptation observed? Steven Peterson, George Hoffer, and Edward Millner examined statistics collected by the state of Virginia about fatal two-car accidents in 1993. In 30 such accidents, one car was equipped with an air-bag, and the other was not. The accident reports identified in each case the supposed "initiator" the party was driving more aggressively. The table indicates that in 22 of the 30 accidents (73%) the initiators drove cars with air-bags, although only 50% of the cars involved in accidents were air-bag equipped. So, the indications are, having an air-bag increased the likelihood of a driver to initiate an accident.

Two-car accidents with/without air-bags (Virginia, 1993)

	With air-bags	Without air-bags
Number of cars	30 (50%)	30 (50%)
Number of "initiators"	22 (73%)	8 (27%)

Source: Adapted from Peterson et al., p. 262.

Any single bit of evidence such as this can only be indicative, not conclusive. Moreover, even if entirely valid, these results do not necessarily imply that regulations requiring air-bags are inadvisable. Although drivers of airbag-equipped vehicles may indeed be more inclined to take risks, the increased aggressiveness may not entirely cancel out the overall beneficial safety effect of air-bags.

This reading material illustrates that air-bags are equipped to reduce the damage of accidents, the intention of which is good. However, the result is out of expectation. Because of the reduction of the damage of a car accident, drivers become aggressive in driving. Under the constraint of the cars equipped with air-bags, the costs of aggressive driving decrease from severe disability or even death to minor injury or disability. As a result, drivers tend to drive faster. The risk of severe damage caused by a car accident is reduced, while that of the happening of car accidents

increases. So the number of victims of accidents may increase. The result is uncertain for drivers, but for pedestrians without air-bags on their backs, they are sure to be harmed. Then for the whole society, air-bags equipped in cars may do more harm than good.

In order to eliminate crimes, some have suggested a death penalty to be imposed on all crimes regardless of their severities. However, once the severity of the penalty was out of line with that of crime, the costs and benefits faced by a criminal would be distorted, and he would make distorted choices. Just imagine what would happen if a nation did do so. The criminal rate would certainly plunge, and minor offences such as stealing would almost disappear, but serious crimes such as murder would increase. Why? Suppose now one was out of money and with no option but to steal or rob. Meanwhile, the police's detection rate cannot be 100%. In other words, one might not be caught and punished. So what kind of crimes would one opt to commit? Will one opt for petty stealing or robbing a bank? The consequences of being caught in both cases would be the same: death penalty. The same costs with different revenues, what would one choose? The answer is obvious. The legislation should base on objective analysis instead of passion, or else there would be horrible consequences. No nation has implemented such a death penalty, which implies it is unfit to survive.

1. This reading material is an excerpt from "Price Theory and Its Application", P9. [C](#)

Lecture 4: Scarcity, Competition, Market, and Non-market

In the last lecture self-interest is established as the postulate of economics. However, we cannot directly use the postulate to explain phenomena, but show how self-interest manipulates people to make different choices under different constraints. There are various kinds of constraints, and in this lecture, we will start with a crucial constraint: scarcity.

Section 1: Scarcity

Scarcity means that compared to people's desire, the goods they want are not enough. Scarcity is an essential constraint because if there were no scarcity, people would not have to make any choice even if they are self-interested. Just imagine that everything we want was unlimitedly satisfying. Thus no matter how self-interested we are, we would always get all of them. So why should we bother to make any choice anymore? However, goods are scarce, and we have to give up something in order to get other things else, so self-interest matters much in carefully calculating how to obtain goods at the lowest price.

In economics, anything of which available is preferred to non-available is defined as a "good" regardless of whether it is a tangible commodity or intangible service, or spiritual enjoyment (including feelings). Some economics textbooks define a "bad" as anything of which non-available is preferred to available, such as garbage and noise. However, with proper redefinition, all "bads" can be converted to "goods", such as garbage to "cleaning up garbage". Thus it is not necessary to analyze "bads," and only "goods" are enough.

If anything of which not only available is preferred to non-available, but also more is preferred to less, it is defined as an "economic good". That means it is scarce, so anyone who wants to obtain it must pay a cost or a price if it is in the market. If anything of which available is preferred to non-available, but more is not preferred to less, it is a "free good" which means it is not scarce and takes no cost.

We have to be cautious of what is called "free" in daily life as it often means no need to pay the price in the market, which is not the same as taking no cost. So what is free in the market may not be a free good. For instance, public space is free to enter, but of which more is preferred to less, and is not a free good.

So what is a free good? Let's take the air as an example. In the place where the air is fresh and clean, of air available is preferred to non-available. Even indispensable, but because it is so abundant that no one will pay a cost for more of it, the air is herein a free good. However, in a big city where the air is often impure, fresh and clean air becomes scarce. People are willing to pay a cost for more fresh and clean air, such as buying air purifiers, paying higher prices for houses with better greening. Air herein becomes an economic good. For underwater workers, the air is so scarce that they will pay a rather high price to buy bottled oxygen. Goods are almost economic goods, while free goods are rare and only exist under certain conditions, which means scarcity is ubiquitous.

Some may envision a future society with extreme material abundance. Can it be a society without scarcity where the distribution of goods could be according to people's want? Let alone whether or not a society in the future can genuinely achieve such extreme material abundance that the supply of some goods is unlimited relative to human's need, even if this could be true, would there be no scarcity?

In colleges, there are always empty seats in classrooms, because number of seats is higher than that of students. Since one student can only take one seat at a time, the supply of seats is unlimited relative to their needs. However, is there no scarcity for seats? Why are more front-row seats occupied while back-row ones mostly left empty? The key is the different qualities of seats! The front-row seats are good, while the back-row seats are bad ones. Thus, although the seats outnumber the students in the classroom, the good ones are still less than the students, so they are still scarce! The so-called good and bad emerge from comparison, so no matter how many seats there are in the classroom, there is only one seat that is the best. Otherwise, how can it be

called "the best"? Similarly, there is also only one seat that is the second-best, and the same to the third, the fourth...best. There is always scarcity!

So this is the key to scarcity: if there is no quality difference among goods, increasing the number of goods can eliminate scarcity. However, the difference in the quality of goods leads to scarcity, even if the quantity of goods far exceeds the want.

Thus, in order to realize the ideal society without scarcity, we need not only extreme material abundance but also eliminate all the quality difference - including the quality difference among human beings. It is a pity that God created human being unfairly. Some persons, e.g. Einstein, are born with higher IQs. Maybe in that ideal society, a person would take an IQ test as he/she is just born, and if his/her IQ were higher than a "standard man," some surgery would be needed to eliminate the higher part of IQ. Also, those whose physical quality is superior, such as sports stars, would need elimination operations. The question is, who would wish to live in such a "Brave New World"?

God created human being both fairly and unfairly. The unfairness has been talked about, while the fairness lies in the fact that no matter how mediocre in many aspects, one can always find a specific aspect superior to others. However, it may not be necessarily useful in the market, so it does not have a price. That is why most people are mediocre. If all the quality differences were to be eliminated, not only Einstein or sports stars would have to undertake the operation, but all of us would do so to eliminate every aspect of quality difference according to a "standard man."

Section 2: Competition, Competition Criterion, and Game rule

Human is self-interested, and the goods he wants is scarce, so there is naturally fight for the goods. Competition means self-interested people fight for scarce goods. There is always competition, as long as there are self-interest and scarcity.¹ The following lectures will talk about the market structure, including free competition and monopoly. However, do not misunderstand that only in the free competition market there is competition while there is not in the monopoly market. There is still competition in the monopoly market, just in another form. So it is totally wrong that economics textbooks or some economists boast about the lack of competition.

In any competition, winners must be selected out of players, but how? Competition in sports is the most obvious one that people can experience, so let us take it as an example. For instance, in a running race, there is a standard or criterion of "speed" to select out the winner: the fastest runner wins, and slow ones lose. In order to make sure that the winner will be the fastest runner, there are a bunch of rules in the game. Such as all players should start from the same line and kick-off at the same time, and drugs are prohibited. Just imagine that if jumping the gun were allowed, the first person crossing the finishing line might not be the one who runs fastest. So the rules of the game are designed to establish the criterion of competition.

Let us generalize the above analysis. Suppose now we go to a grand theatre to enjoy an idol singer's concert instead of taking a lecture in a college classroom. Surely there are more seats in the grand theatre than in the college classroom, but there are even much more people like to enjoy the concert. It is obvious that there is scarcity, so competition is inevitable. How to decide who wins and who loses? How to decide who can get those seats at the concert? For many, the first part of this question may not be easy to answer, but the latter part is rather easy, so almost everyone can answer without thinking, "Tickets for sale!" Bingo! This is precisely the criterion for deciding wins or loses: Price! Whoever is willing to pay the price of the ticket will get the designated seat, while whoever is unwilling will not get one. The essence of using price to decide the winners of the competition is the distribution of scarce goods by a price, which is commonly known as the market economy! In other words, the market economy is the price criterion for competition.

However, in reality, there are many other criteria other than the price criterion for competition.

Take the example of seats in college classroom mentioned above. Although the good seats are still scarce, they are allocated on a first-come-first-served basis instead of paying a price. The early birds occupy the better seat, while the later ones have to sit in back-row. Moreover, the seats may also be allocated by a teacher who has the right to decide which students can sit where, which is the case in elementary and secondary schools. The seats may also be allocated by age or longevity, which is precisely the hierarchical system of seniority. This kind of allocation for seats is prevalent at the meetings of government. The seats may also be allocated by voting, where who gets the most votes can take the best seat. This is so-called democracy.

Last but not least, the seats may also be allocated by force. Although one may arrive first and occupy the best seat, if the latter student is much taller and stronger, and by waving his fist, one may be threatened to give up the best seat to him. This is so-called "the law of the jungle," which is the law of nature, and rules not only human society but also the whole nature.

Criteria like these are countless, all of which are named as "non-pricing criteria" in economics because they all do not use price as the competition criterion.

Why does economics single out the price criterion and leave all other criteria to another category? Because among all the competition criteria, price criterion is the only one that does not lead to the so-called "waste." The word "waste" is not entirely correct, so should be revised to the more strict term "rent dissipation" which is rather complicated, so will be explained in later lectures in detail. Whereas "waste" is commonly known and easy to understand, so it is used here to express the meaning of "rent dissipation" for the time being.

Why does price criterion not cause waste? Because with price criterion, those who want to obtain the scarce goods must pay a price, i.e. paying money. On the one hand, the money the buyers pay becomes the seller's income, which means the money does not disappear in vain from the view of the whole society. On the other hand, where does the money paid by buyers come from? It comes from selling something needed by others in the market. That is to say, in order to satisfy his self-interest, one must benefit others first, to get money to take part in the competition. The more one contributes to others or the society, the more money he gets, and the higher the chance of winning the competition with price criterion. In the last lecture, it is mentioned that self-interest does not necessarily mean harming others, and, in many cases, altruism is the best way to self-interest. The market is precisely a system that guides the self-interested human to benefit others. From the above analysis, we can further understand how the market economy decides winners or losers by price criterion and guides the self-interested human to benefit others first in order to win in a competition.

Someone might doubt: if it is the price criterion that decides winners or losers, wouldn't the rich always win? Firstly, with this competition criterion, the rich are indeed in a more favorable position, and that is exactly why people are motivated to seek more income by benefiting others and creating more wealth for society.

Secondly, even with the price criterion, the rich are just more capable of paying a high price than the poor. It means that they are less limited by the constraint of wealth, but not necessarily more willing to do so. Take the above pop star concert as an example. Let us assume the singer is an idol of the young who are willing to pay hundreds of RMB to buy tickets of the concert, while others are not attracted. The older persons who have worked for years are generally richer than most of the young, but will usually not pay, although they can afford the price. So, in the competition for tickets, the young though poor win, the old though rich lose. If the price of a ticket were low enough, such as a few dozens or even tens of RMB, the old might also be willing to pay. Nevertheless, when the price is too high - higher than what they think to be worth, they will not buy. Similarly, few people will buy high-grade fruit for dozens or even hundreds of RMB. Can't they afford it? Impossible! Even the young who have not worked can afford hundreds of RMB to buy concert tickets for their idol. How come the old cannot afford the high-grade fruit? It is not that they cannot afford, but that they do not think it is worth the price, so will not buy.

Someone might also doubt: may the money people use to bid for goods necessarily come from their previous contributions to society? Maybe they get the money by stealing, robbing or cheating. First of all, in the last lecture, it has been analyzed in details that as long as there are effective laws to punish the crime of robbing, stealing, and cheating, a self-interested human will choose to get money by contributing to society, instead of by harming others. Morality will also condemn such harmful behaviors to increase the non-monetary costs. Secondly, as mentioned above, the criterion to decide winners or losers needs a bunch of game rules to establish it. These laws and morality are exactly the game rules that establish the price criterion!

If a person competes for scarce goods with the money by robbing, is it different from those who win a running race by jumping the gun or doping sneakily? Without the laws that effectively punish those who get money by harming others, a self-interested human will choose to harm instead of benefiting others. It is just as that if the game rules of running race do not forbid

jumping the gun, self-interested runners will choose to try to train the technique of jumping the gun instead of reacting quickly to it.

Furthermore, if robbing is legal, the muggers can not only rob the money for bidding goods but also the goods themselves without paying any price at all. In this extreme circumstance, the price criterion for deciding winners or losers has fallen apart entirely.

In the later lectures, we will explain the legal system to establish the price criterion (i.e. the market economy). It is the system of protecting private property rights or PPRs - the law forbids robbing since it infringes the victim's PPRs. The PPR is a very complex concept which will be explained in the later lectures. As a short introduction: PPRs include use right, the right to income, and the right to transfer. The legal system of protecting PPRs is the cornerstone of the market economy. It is just empty talk if we only talk about the market economy as a competition criterion without legal system as the game rule.

It is said that the market economy should be divided into a good one and bad one. The market economy in developed countries is a good one, while that in the vast majority of developing countries is a bad one although most of them are not socialist countries practicing planned economy. It is implied that in developed countries, the market economy is supported by the rule of law; while in developing countries, there is no rule of law, so they are overwhelmed by bureaucratic corruptions. There are also other different descriptions, such as "crony capitalism". All of these are a severe misunderstanding about the market economy! Without an effective legal system as the cornerstone, there is impossible to be a market economy, because it is not the price criterion but other non-price criteria such as force and power that decide winners or losers of competitions. How can a running race that allows jumping the gun and doping be recognized as a normal one? So, the market economy is not divided into a good one and bad one, but rather into a true one and fake one!

1. Strictly speaking, the condition of "society" must be added, and along with self-interest and scarcity can lead to competition. In a one-man world, no matter how self-interested a person is, he just needs to compete with nature and animals. However, economics as a social science is not concerned with this kind of competition. Economics focus on the competition among humans which only exist in society. The concept of "society" will be defined in the lecture of "transaction costs". [↳](#)

Section 3: The Choice of Competition Criteria

As mentioned above, there is no so-called waste with price criterion to decide winners or losers of competition. However, other non-price criteria will undoubtedly result in waste more or less. For example, first-come-first-served will cause queuing. One stands in line with nothing else to do, and he has paid the cost of time - a kind of opportunity cost that will be explained in the later lecture about costs. While no one gets any benefit from his waiting, so he contributes nothing to society. That is why it is a waste. In contrast, if he pays the price directly, the seller will get the money which comes from the buyer's previous contributions to society.

If it is up to the teacher to decide who gets the best seats, students will try to flatter the teacher, such as licking the shoes or even giving packets of money, which is equivalent to paying the price to the teacher. However, this kind of bribery is prohibited by law. The process of law enforcement (including reporting, filing, taking evidence, trial, and judgment, etc.) involves huge costs and is wasteful. If law enforcement is not strict, teachers will be careless in teaching but obsessed with how to collect packets of money from students skillfully.

If it is seniority that determines who can get good seats, obviously self-interested human will choose to do nothing but wait for getting older, which cause a waste similar to waiting in line. In state-owned enterprises where seniority prevails, there are significantly more "old dogs", which is precisely one of the main reasons for the inefficiency there.

If the allocation of seats is decided by democratic voting, people will form pressure groups to lobby for votes. In the later lecture of "Political Economics," we will explain why it is wasteful, and show that voting on specific issues will result in an infringement of PPRs.

If it is the law of the jungle that prevails, a human will keep strong and spend money on weapons. It is good to be strong, but if it is for harming others, instead of producing or creating more wealth, it is again a waste. Not to mention the weapons are made for massacre.

In conclusion, all non-price criteria of competition are bound to be more or less wasteful, which should be called "rent dissipation" in economic. Rent dissipation is a kind of transaction cost which will be talked about in details in the later lectures of "Transaction Cost" and "Rent Dissipation".

As a smart reader, one might point out: since human is self-interested, and price criterion does not cause rent dissipation, while the non-price criterion surely does, self-interested human should only choose price criterion and never choose any non-price criteria. But this conclusion is obviously not consistent with the fact (i.e., rejected by the fact). The answer is: although price criterion does not cause so-called waste, the market transaction will cause transaction costs, which shall be explained in the later lecture of "Transaction Cost". Also, in order to establish price criterion, there must be an effective legal system to protect PPRs as the game rule whose establishment and operation will also cost a lot, and it is also a kind of transaction cost. Again, the detailed explanation will be in the lecture of "Transaction Cost", and here is only the brief main point. The cost of price criterion or market transaction is the transaction cost in a narrow sense, while the cost of the legal system is that in a broad sense.

Therefore, if price is chosen as competition criterion, there will be transaction cost, while if a non-price criterion is chosen, there will be rent dissipation which is also a kind of transaction cost. So, which will be chosen: price criterion or non-price one? The answer depends on the constraints in a specific circumstance. Under different constraints, the transaction costs involved in price

criterion and non-price criterion will vary. Of course, self-interested human will choose the one that causes the least transaction cost.

Thus, the postulate of self-interest dominates all economic theories, and along with different constraints, determines all explanations of human behaviors.

The name “economics” is in fact somewhat misleading. A layperson would think that economics can only explain the human behaviors in the economic field. However, it is far beyond that field. The behavior of human choice, in the economic or other fields, can be explained by economics. This book has taken numerous examples to illustrate economic explanation which are not limited in the economic field but also involves other fields such as law, political science, history, sociology, and anthropology. From this perspective, the more accurate name for this discipline should be “Science of Human Behaviors”. Some call this as “economics imperialism” which means “invasion” into other fields of discipline by using the logic of economics. More appropriately, economics is the whole of social science, instead of just a part of it. Because social science is the study of human beings, and no social sciences other than economics has found suitable postulates, and base on them to construct a theoretical framework which is logically self-consistent and stand up facts test (i.e., “refutable”). Therefore, only based on the postulate of economics can other social sciences be regarded as “science”, which means they are part of economics, just the applications of economics in different fields.

Furthermore, as mentioned in the last lecture, there is an interdisciplinary “biological economics” nowadays, which extends the postulate of economics to natural sciences. By assuming non-human organisms (such as plants, or even genes) are self-interested, it explains some biological phenomena. Thus, in the broadest sense, economics is the science that can explain all of the following phenomena: by assuming the subjects (human or non-human) are self-interested, regard all phenomena of the subjects as their choice with purpose under certain constraints.

Lecture 5: Positive Economics and Normative Economics

The previous lectures expound that economics is a social science, like physics and other natural sciences, which logically derive a theoretical system based on some postulates and axioms to explain various phenomena in reality. However, as mentioned in Lecture 2, the subject of economics is human, while that of physics is nonliving matter. So economics is severely affected by an economist who himself is also human, and the incompetent or ill-intentioned ones of them often distort the research deliberately or unconsciously and let their personal feelings or even interests interfere with scientific study.

How can economists avoid distorting economic analysis with personal feelings or interests? Moreover, how can the ordinary learn to think like economists even when they do not need to become economist, and free from the delusion by the economists' intentional or unintentional misanalysis? Answer: we should distinguish clearly between positive economics and normative economics.

Section1: Positive economics

Positive economics or “empirical analysis” argues that economic theories are used to explain the phenomena, and answer the question of “to be” instead of “ought to be”. Based on the postulate of self-interest, along with the constraints changes, economics explains various kinds of human behaviors. Economics is not responsible for answering such questions as “Is self-interest good?”, “Should human be self-interested?” or “should a person make such a choice?”

As mentioned in Lecture 2, if there is a one-hundred-dollar bill on the ground, it will be blown away by the wind if there is so, which is a phenomenon that can be explained by physics. However, the bill will still disappear in the absence of wind, which is a phenomenon that physics cannot explain, but economics can. The economic explanation for it is: because human is self-interested, if there is no watching over or surveillance camera, which is the constraint, it is sure that someone will take away the bill. Economics is only responsible for answering why someone will take the bill away, but not answering whether he should take it away. Physics is also only responsible for answering why the bill will disappear when there is wind, but not answering whether the wind should blow it away. However, because physics deals with nonliving matter such as wind, physicists are hardly criticized for not answering the latter question. Whereas because economics deals with human behavior, economists are often criticized harshly for not answering such question as to whether particular human behavior is good or bad, which is even blamed bitterly as immoral!

Imagine that someone threw garbage out of a high-rise building and injured a passer-by. Newton came to answer why the garbage fell to the ground instead of flying up to the sky after leaving human's hand. Smith came to answer why the man threw the garbage out of the building instead of putting it in a trash can. A moralist came to answer whether the man should throw the garbage out of the building.

An economist does not talk about morality, which does not mean he is immoral. As an economist, when he analyzes human behaviors with economic theories, he only needs to answer the question of “to be” instead of “ought to be”. It is morality that is responsible for answering the latter question. It is just the division for different disciplines. Asking an economist to answer whether particular human behavior is good or bad is like asking Newton whether an apple should fall to the ground. To criticize an economist coming to a conclusion you do not like is as ridiculous as to blame an apple which fell from a tree and injured a passer-by, “Damn it! What a bad apple! How can it fell exactly when someone is passing by?”

Section 2: Normative economics

However, government always asks economists questions of "ought to be", such as how should government solve an economic problem, or whether a policy is good or bad and whether it should implement or not.....So how should economists do?

To answer these questions, economists should answer the question of "to be", that is, analyze empirically at first. Take the case of car airbag in Lecture 3 for example. Government might ask economists: "Should be a policy of stipulating all cars to have airbags implemented?" Economists should conduct objective scientific analysis with economics at first, pointing out that because of the self-interested nature of human, with airbags on cars, drivers will drive more recklessly, which lead to more car accidents and casualties. Then economists do not need to say anything more, and government would naturally know how to do. If what government wants are fewer car accidents and casualties, given the economists' analysis, it is clear that a mandatory stipulation of car airbags can not hit the target.

Of course, it is also possible that government has colluded with airbag manufacturers, so what it wants is to force everyone to buy airbags. Obviously it is improper for the airbag manufacturers to openly claim that they push government to implement a policy for their interest, so they come up with a seemingly plausible and grandiose reason. Government explains that airbags can help to reduce casualties when a car accident happens, so they are good for drivers. If the objective and neutral economists present the correct economic analysis mentioned above, the self-interested airbag manufacturers will often deliberately incite the resentments of the public, cause confusions in the argument, and attack the economists for not caring about drivers' life as immoral. The ignorant public cannot objectively analyze the truth, so are easy to cheat. Some of them are drivers, and they wrongly believe that the agents of airbag manufacturers care more about their benefits than the neutral and so-called "immoral" economists. Airbag manufacturers will surely try to hide their real purpose from the public, so they let their agents, including some other economists who appear to have no interest with them, to speak in public. Finally government implements the stipulation about airbags, ostensibly accepting the advice of some economists who are agents of airbag manufacturers while turning his back to the objective and neutral analysis. Some ignorant enjoy the pleasure of scolding the so-called "immoral" economists. Moreover, the drivers have to spend more money on car airbags, not knowing that they may bear a higher risk of car accident casualties!

Such cases happen again and again in reality: government asks for advice from economists about policy, but he may not be for the public's interest. Some economists are either disabled or not objectively neutral. The public can not positively analyze the effect of the policy, so they do not know the truth and enjoy the pleasure of scolding others. The victims are harmed in ignorance, and at the same time, pay for the costs in silence.

The society is full of such self-interested persons who have complicated interests and speak with deceptions. One of the purposes of studying economics is precisely to learn to analyze objectively. Positive economics teaches us not to ask the questions of "ought to be", but only or at least firstly ask the questions of "to be". Please do not believe what the self-interested persons say, but observe what they do. It is not essential what the good intentions of the government's policies are, while what the measurable effects they have mattered because the final effects often run counter with the original intentions.

Therefore, normative economics, or “normative analysis”, must be based on the empirical or positive analysis. Only after the empirical analysis is conducted can we take value judgment into account. Take the example of airbags: what does government want? If he does want to reduce car accidents, it is not necessary to make a policy about airbags. It would be the opposite case for policy-making if what he wants is to benefit airbag manufacturers. A value judgment is not science but ethics, and government often has its idea, so economists should not put a finger in, or else they will just be used for the unspoken purpose.

Section3: Value Judgment

As mentioned at the end of the previous lecture, non-price criteria would lead to rent dissipation which is a kind of transaction costs. Although there is not rent dissipation for price criterion, there is transaction cost (in a narrow sense) for using the market. What is more, in order for price criterion to be established, there must be a legal system protecting PPRs as the game rule, and there is also transaction cost (in a broad sense) for the establishment, operation and maintenance of the legal system. Therefore, whether people choose price criterion or non-price criteria depends on their transaction costs under different constraints. Self-interested people will, of course, choose the criterion with less transaction cost.

In fact, in such cases as to how to allocate seats in a classroom, people will seldom choose price criterion (e.g. auction). However, it is neither decided by democratic voting or by force, but the non-price criterion of first-come-first-served. Why? Because the benefit of the best seat is minor. In order to distribute such a minor benefit, it is not worth taking the trouble to organize an auction or voting, or fight with each other, which means there is transaction cost and it outweighs the benefit. Although there is rent dissipation (a kind of transaction costs) for first-come-first-served as a non-price criterion, it is still the least costly way among all for the allocation of seats in a classroom and chosen by self-interested people.

In reality, there are both market (price criterion) and non-market (non-price criteria) behaviors, but the latter is often ignored in microeconomics (MSE). There is a big issue related to this ignorance: is the market economy better or planned economy better? There was once a heated ideological debate in China, and today it seems that a conclusion has been reached that the market economy is good. Really?

What is the meaning of "good"? If one likes an apple, he will think apples are good. While if one likes pears, he will think pears are good. Here good and bad are subjective value judgments, and there is no objective standard, so it is a question that science cannot answer. However, if an objective criterion (for example, the more the nutrition, the better) is set in advance, and it is also necessary to precisely determine what kind of nutrition it is, who will eat the fruit, and whether he lacks that kind of nutrition, which are all constraints, then science can answer the question "Which is better for someone, apple or pear?" Strictly speaking, science answers the question "which is more nutritious for someone, apple or pear", and further answers the question "should someone eat apple or pear?"

Turn to economics. Is the market economy better or planned economy better? If there is no objective standard set in advance, it can only be judged by subjective value, and economics as a science cannot answer. However, if "the less the cost is, the better" is the objective standard, the question becomes: "is the cost of a market economy or a planned economy less?" Market economy and planned economy are different economic institutions. As will be analyzed in the later lecture on "transaction costs", the cost of an institution is the transaction costs in a broad sense. Therefore, this question is actually about the transaction costs of the market economy and planned economy. However, it can be seen from the above example of allocating the seats in a classroom that it is not enough when there is only objective standard because specific constraints are also needed. Without constraints, it is impossible to identify the transaction costs of different competition criteria.

As mentioned in the previous section, normative economics must be based on empirical analysis, along with objective standard which should be set in advance. Firstly in empirical analysis, different choices are analyzed under specific constraints. Then according to the objective standard set in advance, economics as science can answer the questions of "ought to be."

Section 4: The Problems of Welfare Economics

In economics, the normative analysis is concentrated in a branch called "welfare economics". Do not misunderstand that it is the study of the welfare economy; instead, it tries to use "waste" or "inefficiency" as an objective criterion to answer the question of "to be".

Welfare economics will be discussed in the later lecture of "The Coase Theorem", and herein we just give a brief. There cannot be "waste" or "inefficiency" in economics, which is a direct conflict with the postulate of self-interest. In Lecture 3, it has been emphasized that there cannot be any theory or explanation conflict with the postulate of self-interest. In that lecture, it has also been pointed out that the so-called "irrationality" is wrong in that information costs (a kind of transaction costs) are ignored, so human is rational to choose to be ignorant. The so-called "waste" or "inefficiency" in welfare economics is also wrong in neglecting transaction costs.

In Lecture 1, it has been cautious that there are many mistakes in MSE. They stem from two primary sources, one of which is the omission of critical constraints, and transaction costs are precisely the constraints that are often overlooked. The theories neglecting transaction costs will be a conflict with the postulate of self-interest, and of course they are serious mistakes, just as the laws that are conflict with the Constitution are serious mistakes. If a law is unconstitutional, it must be amended to conform to the Constitution, or even repealed. According to the same logic, welfare economics, which takes "waste" or "inefficiency" as the criterion of value judgment and is a conflict with the postulate of self-interest, should be abandoned.

Lecture 6: The History of Economic Thinking -- Microeconomics and Macroeconomics

Economics is divided into microeconomics and macroeconomics, which copies the division of physics. Economics has copied many practices in physics which we will meet in the later study. The division of micro and macro is just one of them.

There is macro-physics first, then micro later. Newtonian mechanics belong to macro-physics which is measured in meter. And micro-physics study molecular and atomic levels of matter where the laws of movements are different from those followed by macroscopic objects.

Unlike physics, there is microeconomics first, then macro later. Why is economics also divided into micro and macro? We need to trace back the history of economic thinking.

Section 1: From Classical Economics to Neoclassical Economics

Adam Smith's "Wealth of Nations", published in 1776, marks that economics has become an independent discipline. From then on to today, economics has gone through two stages: classical economics and neoclassical economics. Until today, it is still in the neoclassical period. In addition to Smith, the famous classical economists include David Ricardo (1772-1823), John Stuart Mill (1806-1873), etc. Alfred Marshall (1842-1924) published "The Principles of Economics" in 1890, which marked the transition from classical economics to neoclassical economics, so he is the second greatest master only next to Smith in economics. His "Principles of Economics" is regarded as the first textbook in economics, and today's economics textbooks all follow it in the structure. The masterpieces of classical economics such as Smith's "Wealth of Nations", Ricardo's "On the Principles of Political Economy and Taxation" and Mill's "Principles of Political Economics" are obviously different from the current economics textbooks which are quite similar to Marshall's "Principles of Economics".

There were no terms of micro and macro in classical economics, but there were the concepts of them which were very different from those in neoclassical economics. In classical economics, there are two fields of study: resource allocation and income distribution. The former could be regarded as micro and the latter to be macro.

Classical economics focuses on analyzing how the market mechanism (price criterion) guides the allocation of resources for production, and how the income from the sales of products is distributed to the owners of the resources. Classical economists advocated the free market and believed that government only needed to play the role of "a night watchman". There are two duties for a night watchman: one is to tell the time, the other is to patrol for fire prevention and security. In other words, in the view of classical economists, the role of government was minimal, and it should not interfere in the market. The spontaneous working of the market was enough to arrange all kinds of economic activities in a society correctly.

Even transmitting to neoclassical economics, economists were still of this opinion about the free market, and only introduced mathematics (mainly geometry) to present and deduce economic theories. Economists had long believed in the market.

Section 2: The Rise and Decline of Keynesian Economics

However, the Great Depression (GD) triggered by the stock crash in the United States from 1929 to 1933 which swept through the significant economies shook the public's trust in economists, thus shaking economists' trust in the market and even the economics itself.

According to the theory of supply and demand that will be introduced later, the market can adjust demand and supply through price changes so that they are always equal. However, it seemed that the orthodox theories in economics could not explain the phenomena of overproduction (i.e. the supply of products exceeds the demand), continuous price decline (deflation, i.e. the demand of money exceeds the supply) and high unemployment rate (i.e. the supply of labor exceeds the demand) during the GD.

Especially the "Say's Law" which believed that "supply would automatically create demand equal to itself" was regarded as an infallible law at that time. From the perspective of it, supply and demand could not be out of balance, let alone for a long time.

During the GD, there was a bitter joke: A son asked his father, "Why don't we buy coal for warming during this cold winter?" The father replied, "Because I lost my job, we have no money to buy coal." The son asked again, "Why did you lose your job?" The father replied, "Because the coal company I worked for could not sell out the coals and went bankrupt, we as the employees were all fired."

Here is what the history textbooks record: President Roosevelt of the US stood up at this critical moment to implement the "New Deal" and turned the tide by government interventions in the market, saving the US economy, ending the GD, also ending the history of letting the free market spontaneously adjust the national economy, and starting the era of big government.

Some views of Keynesian economics have been taught in the ideological and political education of middle school in China. It is the story in textbooks: British economist John Maynard Keynes (1883-1946) wrote his "The General Theory of Employment, Interest and Money" in which he criticized the orthodox economics that advocated the free market, and then established a set of theories analyzing the national economy and advocating government interventions in the market.

Unfortunately, these so-called "historical facts" which are vividly described in textbooks are wrong because of severe misunderstanding and misinterpretation. Friedman, the founder of the monetary school, co-authored with Anna Jacobson Schwartz (1915-2012) "A Monetary History of the United States, 1867-1960". With a large number of facts, excellent statistical skills and dominant monetary theory, Friedman discovered the truth of this period of history by strict logical economic analysis, completely overturning previous understanding of the causes of the GD.

Roosevelt's New Deal was implemented earlier than Keynes's "The General Theory". On the contrary, Keynes was inspired to write his book by Roosevelt's New Deal.

However, it was not Roosevelt's New Deal that freed the US from the GD, but war! The unemployment rate is a definite proof. At the beginning of the GD, the unemployment rate in the US was as high as 25%, which means one in four was unemployed. Later, the rate was reduced to about 10% due to the automatic adjustment of the market and the public works of the New Deal, but it was still severe unemployment. It was not until after the US went to the World War II that the problem of high unemployment had been really solved because many people joined the army to fight overseas, which significantly reduced the supply of labor, and the increasing demand for the war materials boosted the demand for labor. Only then did the US truly get rid of the GD. It

was in 1929 that the stock market crashed, and it was in 1941 after the attack on Pearl Harbor that the US went to war, while it was since 1933 that Roosevelt's New Deal had been implemented when he became president, so there are nearly ten years between them. If the New Deal was effective, the problem of high unemployment rate should have been solved.

Besides Roosevelt tried to get rid of the GD with New Deal, Hitler also led the Germans to another road of self-rescue by starting a war, extricated Germany from the crisis much faster than the US, and even dominated the European continent, of course just temporarily.

There is an idiom in China called "quenching thirst with poison". If it is judged whether it is useful only from whether it can quench thirst (get rid of the GD), poison (Hitler's militarism) also works, but the final result is death. However, Keynes's prescription is also a chronic poison, and just because its toxicity cannot be observed immediately, it seemed at the beginning to be effective, just as when Hitler came to power, he seemed to save Germany out of crisis overnight. So Keynesian economics was thought to be the remedy of economic diseases caused by market failure, and was widely used in almost all economies after World War II.

Keynes was so ambitious that he originally intended to replace the traditional economics since Adam Smith completely. However, some economists made a compromise that traditional economics was also right, except that the scope of application was different from Keynesian economics. The former focuses on how the micro-entities in the market such as individuals and enterprises consume, produce and exchange guided by price. And the latter is concerned about how government closely monitors the national economy, just like a doctor caring for patients, and prescribes according to Keynesian economics to regulate the economy, ensuring that it moves forward on a healthy track. So these economists separate traditional economics from Keynesian economics, arguing that they apply to different fields, just like physics.

Thus, the traditional economics is renamed as "microeconomics", and Keynesian economics is called "macroeconomics". By the way, it was Keynes's father who first introduced the concepts of micro and macro into economics from physics. This compromise of the traditional economics and Keynesian economics is known as "Neoclassical Synthesis" which includes both market mechanisms advocated by neoclassical economics and government regulations or interventions in the market by Keynesian economics. However, Keynes himself did not agree with neoclassical synthesis. Even Keynes and Keynesian School are of different opinions. In Keynesian School, neoclassical factors are added into Keynesian economics. For example, the IS-LM Model is established by this school instead of Keynes.

However, in the 1970s, the sequela of governments taking Keynes's prescription for long terms finally broke out. If the end of World War II in 1945 is taken as the starting point, it just took less than 30 years. Seen from the angle of long history, it was not a very chronic poisoning. It should be the oil crisis in the 1970s that accelerated the outbreak of the sequela, and a new term "stagflation" spread quickly, which was a phenomenon that Keynesian economics could not explain. In Keynesian economics, inflation and economic stagnation should replace each other, and could occur at the same time.

There have been questions on Keynes in economics, which was just not noticed by the public before. The Monetary School led by Friedman came first, followed by the Supply School and the Rational Expectation School. The neoclassical synthesis was also criticized to divide economics into micro and macro with entirely different application scope and irrelevant to each other. Some economists advocate "supplementation of the micro foundation for macro" which means that macroeconomics should be rebuilt based on microeconomics or neoclassical economics. However, it is more thorough that economics should not be divided into micro and macro at all!

Section 3: Price Theory and Monetary Theory

In some universities, such as the University of Chicago and the University of California, Los Angeles (UCLA) in the US, the appellations of micro and macro are considered misleading, and the terms of "price theory" and "monetary theory" are respectively used to replace them.

As mentioned above, economists of neo-classical synthesis classify traditional economics as microeconomics which focuses on the consumption, production and exchange of the micro-entities (individuals and firms) in the market guided by price. So government intervention is absent in microeconomics, or even if there is a government, it only behaves either like a consumer buying various goods in the market or a producer providing public services to the market. However, in macroeconomics, the role of government is exceptional. It functions as the regulator of the economy (mainly referring to the market), which is entirely different from the role of individuals and firms.

Those who use the terms "price theory" and "monetary theory" instead of micro and macro are concerned about another difference that they think is essential: money.

Microeconomics studies how the price mechanism (price criterion) plays a role in the market and guides different economic entities to make choices. Here the price refers to the relative price in barter where there is no money, or there is money, but the value of money is stable, and the price level is also stable (i.e. there is no inflation or deflation), so the nominal price (monetary price) of a good is the same as its real price (relative price). Thus, there is no monetary effect on the economy, which means "monetary neutrality" advocated by classical economics.

By contrast, macroeconomics focuses on money. There is not only money in the economy, but also the value of money may fluctuate, which means the price level is unstable (i.e. there is inflation or deflation). Then the monetary price of a good is different from the real price, so affecting (actually misleading) the economic entities' choices. And the money in modern nations is all issued by government: most are national governments, few are foreign governments, and Euro is the most special, issued by the European Central Bank, a super-government institution.

Therefore, the source of instability of money value directly stems from government which also plays an exceptional role. However it should not interfere in the market and regulate the economy, but should control the money supply to stabilize the value of money and the price level, providing a stable platform for the automatic adjustment of the market.

Obviously, it is monetarists who advocate to replace micro and macro with price theory and monetary theory.

Section 4: The Causes of the Great Depression

In the previous section, we have introduced briefly the history of economics from the classical period to the neoclassical period and how it “split” into micro and macro. Next, the causes of the Great Depression will be explained to illustrate why Keynesian economics was not the proper remedy to macro-economy, but just a chronic poison.

It is well known that the Great Depression began with the crash of the United States stock market in 1929. However, it is not well known that there were stock crashes more severe than that of 1929 both before and after the GD, but they all did not cause such disastrous consequences.

Some may argue that there were not many investors in the stock market before 1929, so the severe stock crashes occurred in the past did not have so massive impacts as that of 1929. However, after 1929, there are more and more investors in the stock market. Why have more serious stock crashes not brought about as severe consequences as the GD? They may argue that after 1929, governments had Keynes's prescription to prevent the spread of the stock crashes to macro-economy by intervention with the market. However, during the colonial period in Hong Kong, British-Hong Kong (BHK) government had been implementing the policy of “positive non-intervention” until it was terminated by the last Governor Chris Patten. Hong Kong was the last piece of land on the earth where Smith's classical thought of “government only needs to play the role of night watchman” was still followed in modern society. During this period, Hong Kong had experienced some stock crashes not less severe than that of the US in 1929. Moreover, the BHK government had never taken any Keynesian prescriptions, but Hong Kong survived unscathed again and again, and the impacts of the stock crashes had never spread to macro-economy of Hong Kong and turned into economic crises.

Why? The answer is also the truth of the GD. The truth is: the cause of the stock crash in the US in 1929 was utterly different from that of most stock crashes. The real cause of that stock crash was also the real cause of the economic crisis which the stock crash later turned into. In other words, the stock crash is just the prelude to the GD, not the cause of it. A stock crash alone will not necessarily cause an economic crisis. However, neither will an economic crisis necessarily turn into a great depression. It was a series of mishandling by the US government that turned an ordinary economic crisis that had happened many times before into an unprecedented economic catastrophe - the Great Depression!

Friedman in his “A Monetary History of the United States, 1867-1960” traced back in detail the data of money supply before and after the stock crash in 1929, eloquently showing that before 1929, the US central bank (the Federal Reserve, or the Fed) injected a large amount of money into the economy. The later lecture of “Monetary Theory” in this book will expound that if the growth of the money supply is out of line with that of the real economy, there will be inflation. However, people (even economists) at that time knew little about monetary theory. That is why the truth or real cause of the GD has been misunderstood for long.

When producers found that the price (nominal price) of goods was rising, they would think that their products were so popular in the market that the demand exceeded the supply. With the stimulus of rising price, they significantly increased production and employed more workers. Because the demand for labor increased, with the limited labor supply, the wage also increased. Workers found that their wages increased, and producers found their profits increased. Everyone found his income increased, and all spent more, so in the short term, the increased supply seemed to be consumed by the increased demand. Investors also spent more money in the stock

market, firstly because more income would naturally lead to more savings (the so-called "investment" in the stock market is the same as savings in economics), and secondly because the increase in profits of enterprises made investors full of confidence in the rise of stock prices. As long as there are enough investors who had such confidence, and more chose to buy in stocks instead of selling out, the stock prices would keep rising.

However, everything had an end. It was too much money flowing into the economy that caused "Money Illusion" in the short term, which made people not know that the general rise in prices was inflation, not real economic growth. The illusion is after all only an illusion, not reality. What was praised by the historians as the "roaring 1920s" was just an illusion of prosperity stimulated by inflation in the eyes of the economists. Sooner or later, producers would find that everything was more expensive, not just their products, and their costs also increased, squeezing their profits, so they no longer expanded the productions, but shrank. Consumers also would find that everything was more expensive, not just their incomes, so they cut back on spending. All of a sudden, the wind changed, and the stock market is the most sensitive to the change of economic wind, so the stock price suddenly turned to plummet. Previously, due to the soaring stock prices, a large number of funds that could not bear high risks (such as bank deposits) were attracted into the stock market. The sudden collapse of stock prices wiped out these funds overnight. As a result, several banks went bankrupt, triggering a crisis of confidence in banks. Even the banks that had no problems in trouble were drawn into a run on the banks.

These problems in finance quickly spread to the real economy where there was overproduction caused by the illusion of inflation. It is necessary to realize that the overproduction was not the fault of the market at all! It was the central bank (the Fed) that issued too much money, which led to inflation and misled producers into thinking that the rise in product prices was due to the rise in demand, but it was wrong. What had risen was only money supply, so the nominal or monetary price of goods increased, but the real or relative price of goods had not changed at all.

Nevertheless, how could producers know that? In a monetary society, what producers could directly observe is the monetary price of their products, and could not know if the central bank had issued too many money, so they would respond to the price rising by expanding productions. If the price as a signal for the demand and supply had been distorted by inflation, the producers' response would be wrong, because the increased supply had no corresponding increase in demand to consume. That is precisely the real reason why there is a so-called overproduction.

The traditional economics (the so-called microeconomics) thinks that the market will pass information about supply and demand to various economic entities by the rise and fall of prices. Prices will rise when demand exceeds supply and fall when demand exceeds supply. All economic entities are self-interested or rational to respond accordingly: when prices rise, producers increase supply and consumers reduce demand; while when prices fall, producers reduce supply and consumers increase demand. These responses are all pushing the changes of supply and demand to be equal. Therefore, it is inconceivable that there will be overproduction for traditional economics. However, the problem lies in the fact that there is money in reality, and the central bank may implement wrong monetary policies due to ignorance or intentionally – before the GD, the Fed were ignorant, but most of the inflations were initiated intentionally. The wrong monetary policy will lead to instability of the money value which appears as the monetary price out of line with the real one. Thus, the information implied by the money prices is not only about supply and demand, but also mixed with the "noise" of inflation, which misleads producers (in fact, including consumers) to respond not to push the changes of supply and demand to be equal, but to be unequal.

Is this the fault of market or government (central bank)? Who has distorted the price? Who has made the information implied by monetary price mixed with the noise of inflation? Who has misled producers and consumers?

There was an even more terrible story to come. The inequality of supply and demand caused the stock crash, which caused a bank crisis. The sharp contraction of the money supply caused by the bank crisis will, in turn, worsen the economic crisis, which will be explained in the later lecture of "Monetary Theory". In a word, according to the monetary theory understood only by later economists, central bank should increase the money supply when the economy was in recession. However, the Fed did not realize that it had made a mistake of injecting too much money into the economy until after the stock crash, and suddenly changed the monetary policy to tighten the money supply, which is precisely the opposite of what they should do in a financial crisis. It was so dramatic, and it should be said to be tragic! In order to correct a mistake made before, a more significant mistake had been made at the wrong time.

As a result, the US economy worsened from an economic crisis characterized by overproduction which had happened before, to a great depression characterized by a continuous sharp fall in prices (severe deflation). The price could not even make up for the cost of transporting the products to the market, so producers chose to destroy them all on the spot to reduce the loss. That was the reason why there were the ridiculous phenomena recorded in history textbooks: milk was poured into the river, and chickens were drowned in the water, instead of being sold out.

Section 5: More Truths about the Great Depression

The cause of the Great Depression explained by Friedman's "A Monetary History of the United States, 1867-1960" has been expounded in the previous section. However, this terrible story had more tragic developments, because Friedman only explained the phenomena of overproduction and severe deflation, but did not explain why there had been such a high unemployment rate for ten years. The authorities of the US, not just the Fed, had made a series of serious mistakes, which coincidentally joint together to cause such an unprecedented GD.

The first big blunder was that the US Congress had just approved business income tax which was vehemently opposed by the traditional economists because it was suspected as double taxation. The profits distributed to the corporate owners after the business income tax would be taxed again by the personal income tax. Here we would not discuss if the business income tax was true or false, but point out that it was equivalent to tax increase. Even Keynesian economics would suggest that in an economic recession government should stimulate the growth of the economy with a tax cut. However, the US Congress at that time increased instead of a decreased tax, which of course further aggravated the shutdown of enterprises, thus increasing unemployment.

The second big blunder was that the US Congress was so foolish to approve the infamous Smoot-Hawley Tariff Act in 1930 to substantially raise import duties despite strong oppositions from other countries. Different from the US today, it had trade surplus at that time, how stupid it is to embrace protectionism with such "favorable" trade balance! It was easy to guess what would other countries respond. Yes, they retaliated by also raising the tariffs on the exports of US. As a result, the tariff war broke out, and the international trade plummeted to only one-third of the original value within one year. The hard-hit exports naturally could not help the US economy to get rid of the GD as the Tariff Act had expected, but just made it worsen on the contrary.

The third big blunder was the biggest one, whose influence was much worse than the mistakes made by the Fed and the US Congress. It was the minimum wage, trade unions and other institutions that hinder the wage reductions, which led to high unemployment for ten years. Even with chaotic monetary policies which caused severe deflation, and domestic and foreign tax increase, there would not be large-scale and long-term unemployment so long as prices (wages) in the labor market could be flexibly adjusted downward.

Firstly, the minimum wage, trade unions and other similar institutions nullify the essential role of the (labor) market in equalizing the supply and the demand by adjusting the prices (wages). In the economic recession, the productions shrink, so the demand for labor decreases. The market is supposed to equalize the supply and the demand by adjusting the price downward. However, if the equilibrium wage which can equalize the supply and the demand in labor market to eliminate unemployment is lower than that stipulated by minimum wage, the market's adjustment function will, of course, fail to work. If the supply of products exceeds the demand, they cannot be sold; if the supply of labor exceeds the demand, they cannot be sold either, which is precisely unemployment. Is it the market's fault, or the government's? Who forbids the market's adjustment function to work? Who causes the labor market to be disequilibrium with oversupply? Who forces those who may be willing to sell their labor at a price lower than the minimum wage to become unemployed without any income?

Secondly, the profits of enterprises do not necessarily decrease due to price fall in the product market (deflation) and tax increase, which have pushed many enterprises to the edge of a shutdown. However, if the labor cost can be reduced along with the profits, the burden of the enterprises can always be lightened to some extent, thus saving some of them to survive. The survival of enterprises is the most reliable guarantee for the employment of labor.

Reviewing the truth of the GD, one can only sigh and conclude that how wronged traditional economics (and economists) are! However, the most seriously wronged is the market! The market has done nothing wrong. It is human who distorts the price signals with wrong monetary policies. It is also human who with minimum wage law, trade unions and other similar institutions nullify the market's adjustment function of equalizing the supply and the demand. However, again it is human (such as Keynes) who blame the failure on the market, claiming that if the market is allowed to operate freely, it will cause insufficient aggregate demand, overproduction, high unemployment etc. So the government needs to be like a hero in an economic crisis, standing up to correct market's errors. It is similar to that someone buys a machine and does not follow the instructions of the manual to mishandle it, causing such a severe problem that others are hurt or even killed, but he blames the machine for having design defects!

Well, the market is not a person, so it can not stand up to defend itself against human's slander and false accusation. It does not matter, because it is the economic rule that is the final judge. One can wrong the market, but he cannot escape the punishment of economic rule, just as one is sure to fall from the sky and lose his life if he does not follow Newton's laws. As the history of Keynesian economics mentioned in section 2, only 30 years later, the punishment of economic rule had come.

Until nowadays, there have been such punishments in various countries around the world. The specific plots of tragedies may be different, but there is no fundamental difference in essence. The latest ones are the financial turmoil in the US in 2008 and the sovereign debt crisis in Europe in 2009.

Human should keep his eyes open to see the truth clearly: it is human, not the market, who has been torturing himself!

Lecture 7: Equilibrium•Optimum•Margin

Section 1: Equilibrium and Optimum

As mentioned in the previous lecture, economics has copied many concepts from physics. One case is the division of micro and macro, while here is another case of the concept of "Equilibrium" which is very important in economics.

Equilibrium in physics means that an object is in the state of static or uniform linear motion due to the force balance. For example, the mobile phone on the table is static. Why? Because it is now subject to two forces, one is gravity, and the other is the force of the table supporting it. The two forces are of the same magnitude, but of the opposite directions, so reach a balance, and the mobile phone is in a static state.

Economics has copied the concept of equilibrium from physics which refers to a state of static (there cannot be the so-called state of uniform linear motion) when the economic forces are of the same magnitude and the opposite directions. For example, the theory of supply and demand analyzes the two most conventional economic forces in the market: the buyer's demand for commodities and the seller's supply of them. These two forces work in opposite directions: buyer's demand pushes the price to rise, while the seller's supply pushes it to fall. When they are of the same magnitude, the price will stabilize and remain stationary.

However, economists later find that there are cases when there cannot be at least two economic forces to reach a balance. After all, this concept is copied from physics, and, inevitably, there will often be cases that are incompatible with it. Therefore, economists have revised equilibrium to refer to the state that under certain constraints, each economic entity will no longer adjust his behavior (i.e. remain stationary). Why do economic entities no longer adjust their behaviors? It must be because, under such constraints, they have reached the most favorable positions for themselves, so they are satisfied with staying there, not changing any more. If they have not reached the most favorable positions, the postulate of self-interest implies that they will continue to adjust their behaviors. Equilibrium used in consumer theory, the theory of firm and game theory is such meaning instead of the "force balance" directly copied from physics.

In economics, the concept of "optimum" is also often used to describe the above state that individuals adjust their behaviors according to the constraints to make themselves in the most favorable positions. Optimum and equilibrium are related to each other because equilibrium is the overall result when all economic entities have chosen the optimal behaviors. In other words, it is optimum from the perspective of an individual, while it is equilibrium from the perspective of all as a whole.

However, equilibrium in economics is problematic, since this concept in physics is a fact, while it is not in economics. Look out of the window. Cars are running roughly the road or parking on the side, and which one is in the state of equilibrium in physics? The answer is obvious: the parked cars and the running ones in uniform linear motion which can be measured with appropriate tools. However, look out of the window, is the world in the state of equilibrium in economics? No matter the answer is yes or no, what is the reason? Equilibrium in physics is a fact which is observable or at least measurable with tools, while equilibrium in economics is not a fact because it is neither observable nor measurable.

Another criticism that can give a fatal blow to this concept is that since equilibrium refers to the state when individuals all reach the optimum under certain constraints, how can there be disequilibrium? Disequilibrium should mean that individuals do not reach the optimum under certain constraints, which is a conflict with the postulate of self-interest. How can a self-interested person not reach the optimum?

Some may argue if the constraints change suddenly, the individuals who were in the original state of equilibrium will not have time to adjust their behaviors, so they are in the state of

disequilibrium before they adjust their behaviors to reach the new equilibrium. However, it only means that there is transaction cost to adjust the behavior (including the information cost needed to realize that the constraints of the original equilibrium have changed), which is also one kind of constraints. It is wrong to think that individuals are in disequilibrium before they adjust their behaviors to reach the new equilibrium in that the constraint of transaction cost is neglected. In lecture 5, it has been warned that there are many mistakes in mainstream economics, one of which is precisely the neglect of the transaction cost as a constraint. Once the constraint of transaction cost is neglected, there will be mistakes conflict with the postulate of self-interest. The problematic concept of "disequilibrium" once again proves this warning.

In short, if all the constraints are taken into account, there is impossible to be disequilibrium in economics. Then what is the sense of equilibrium? Some economists, such as Coase and Steven N. S. Cheung, have suggested altogether abolishing this concept. However, equilibrium has been used too long and too widely in economics, and intertwined with other economic theories and deeply rooted, so the cost of completely abolishing it is too high. I suggest retaining this concept, but all disequilibrium analysis must be swept out of economics because they are conflict with the postulate of self-interest. Thus the answer to the question whether the world is in the state of equilibrium is: always yes! Anyone who answers no is equivalent to claiming that he does not accept the postulate of self-interest.

With transaction cost, some seeming disequilibrium cases are equilibrium ones, which is quite similar to the cases with friction in physics. Imagine a heavy object is pushed by force, but stay still or in equilibrium. Why is that so? It is because the object is also subjected to a friction force in the opposite direction to the push, and of the same magnitude to it, so they offset each other, and the object remains stationary. However, as long as the push to the object is large enough, it will eventually be a move against the friction.

Similarly, according to the TSD in economics, if the supply or the demand has changed, the original equilibrium price should fluctuate. However, in reality, price will not change every minute with the always-changing supply and demand. Is the price in the state of disequilibrium? Have facts refuted the TSD? There are cases where the price changes every minute with the always-changing supply and demand in the stock market. It is the transaction cost of adjusting the price every minute that prevents a self-interested human from doing so unless the increase in income from adjusting the price is higher than the transaction cost. The stock prices fluctuate every minute because the stock market is precisely the place with the lowest transaction cost for price adjustment, where a complicated computer system is used to match the supply and demand automatically.

If the TSD can only be used to explain the stock price changes, how narrow its application is! In order to avoid falsification of the TSD, it needs to be supplemented with the constraint of transaction cost in price adjustment. With this constraint, a theory with much stronger explanatory power can be deduced, giving new impetus to the traditional TDS: the higher the transaction cost of adjusting price, the lower the frequency of price fluctuations, vice versa.

For example, vegetables are sold both in farmers markets and supermarkets. The prices of the vegetables in farmers markets are more flexible than in supermarkets because the transaction cost for price adjustment of the former is lower than that of the latter. The farmers are the owners of the vegetables they sold and have the complete right to set the prices. By contrast, salesmen in the supermarkets are not vegetable owners and have to report to their managers to obtain permissions if the prices need to be adjusted. It is evident that the internal decision-making process of a supermarket is much more time-consuming, so it is possible that when the day is over and the vegetables have been rotted away, the decision about price adjustment has not been made. The much more time-consuming decision-making process implies the much higher transaction cost in supermarkets than in farmers markets, which is the reason why the prices of the vegetables in the latter fluctuate more frequently than in the former.

Thus, the world is always in a state of equilibrium. The so-called disequilibrium phenomenon which seems to refute the TSD is merely due to the neglect of the transaction cost in adjusting prices. It is not a tautology; otherwise, it is also tautology that friction is used to explain why a heavy object may not be pushed to move in physics. More importantly, the above example of vegetable sales demonstrates how to identify different transaction cost by observing different phenomena. Based on these observable facts, instead of an unobservable concept like equilibrium, we can explain why the vegetables in farmers markets and supermarkets have different frequencies of price fluctuation, which rejected the statement that the vegetables in supermarkets with lower frequencies of price fluctuation are in a state of disequilibrium.

Therefore, economics only needs Static Analysis (the analysis of a particular equilibrium) and Comparative Static Analysis (the analysis of comparing two different equilibriums) which is the analysis of the difference in the constraints of the equilibriums. What is more, economics needs neither disequilibrium analysis nor Dynamic Analysis (the analysis of the process of changing from one equilibrium to another, and it is disequilibrium in the process), because they are wrong in neglecting transaction cost.

Besides, there is another reason why I suggest retaining the concept of "equilibrium". There cannot be disequilibrium analysis, but disequilibrium can be used as a tool to help to check if a theory or explanation is right.

For example, the conclusions of any economic theories must not be in conflict with the postulate of self-interest, which means they must be in the state of equilibrium. So if the conclusions deduced from a specific theory imply disequilibrium, it must be wrong, the logic of which is equivalent to the reductio ad absurdum in mathematics. It is quite useful to judge whether a theory is wrong, so it is a use of disequilibrium in testing theories.

Furthermore, if there is disequilibrium in the explanation of a phenomenon, some constraints (often transaction cost) must have been neglected. As far as theories are concerned, economics is not as difficult as natural science like physics, chemistry and biology. However, the difficulty of economics lies in the applications, and mainly lies in investigating the constraints. There may be thousands of constraints relevant to a phenomenon, so it is not easy to screen out the key constraints, which means that information cost is an important constraint in the study of economics. Here disequilibrium helps to find out whether some key constraints are missing. However, it cannot directly tell what the specific constraints are, which still depends on investigations and perhaps experiences. After all, it is still a handy tool to detect omission in an explanation.

In brief, disequilibrium can be used as a tool to check if the conclusions of a theory are in conflict with the postulate of self-interest and if there are omissions of constraints in an explanation.

Section 2: The Concept of Margin

What will be introduced next is the analysis tool for finding an optimum which is the essential analysis tool in economics: marginal analysis.

In Lecture 3, the postulate of self-interest has been explained in detail, which refers to that human will try to achieve the best result with the least cost. It is the same meaning of optimum.

As mentioned in Lecture 6 about the history of economic thinking, an essential development in the process from classical economics to neoclassical economics is the introduction of mathematics as a deductive tool. Then how to interpret the postulate of self-interest in terms of mathematics? The answer is the method of finding extrema: the least cost is finding minimum value, while the best result is finding maximum value. Furthermore, how to find extrema in mathematics? In geometry, it is to find where the slope of a curve is zero, or the tangent of a curve at that point is horizontal. While in algebra, it is to find when the derivative is zero.

A new and important concept of margin is born by the introduction of this mathematical thought (not only the calculation) into economics. From the 1870s to the beginning of the 20th century, the British economist William Stanley Jevons, the French economist Léon Walras of Lausanne School and Carl Menger of the Austrian School independently put forward "margin", which is called "marginal revolution" in the history of economic thinking. From then on, there is marginal analysis everywhere in economics.

What is "margin"? Since this concept initially came from mathematics, the following expounding will start from the perspective of mathematics, and then extend to the economic implication which is not entirely the same as mathematics.

Readers should be very familiar with two kinds of magnitudes: total magnitude and average magnitude. For example, the total weight of all people in a room is a total magnitude, while the average weight or the weight per capita is calculated by dividing the total weight by the number of people. However, the most crucial magnitude in economics is not total and average magnitudes but marginal magnitude. If someone walks in, which increases the number of people, the total weight will rise. The weight of that person is a marginal magnitude. Moreover, if someone walks out, which decrease the number of people, the total weight will fall. The weight of that person is again a marginal magnitude.

Let us take an economic example. The two geometric diagrams in Fig.7-1 respectively show the relationship between total revenue (R), average revenue (AR) and marginal revenue (MR). The upper diagram pictures the R function, and the lower diagram the associated AR and MR function. Warning: A total magnitude (such as R in the upper panel) should never be plotted on the same diagram as average and marginal magnitudes (such as AR and MR in the lower panel) because the units of measurement are not the same. The vertical axis of the upper diagram of Fig.7-1 is scaled in dollars, whereas the lower diagram is scaled in dollars per unit quantity.

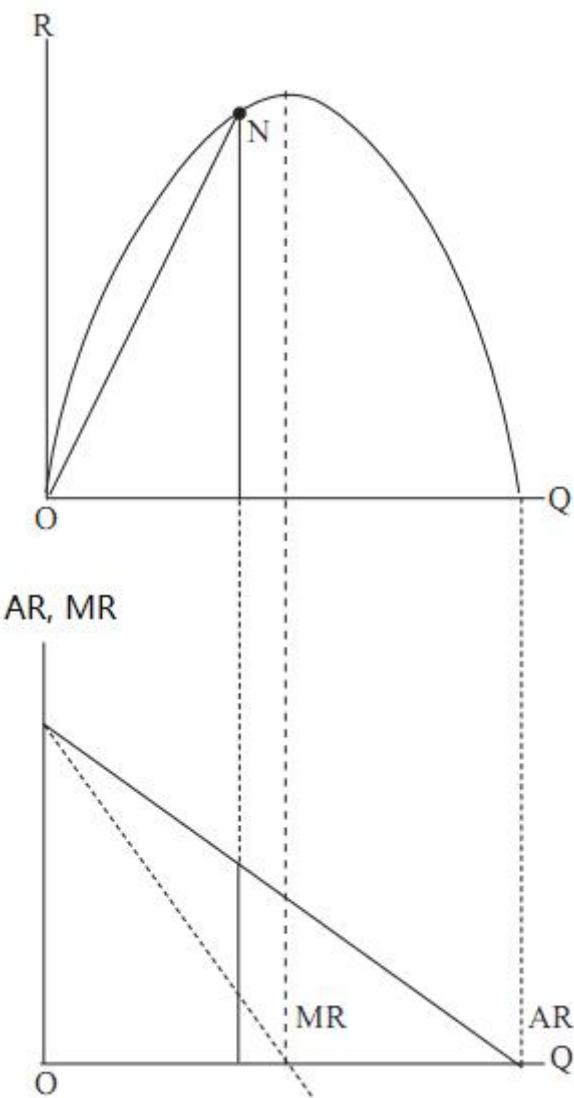


Figure 7.1

In algebra, the average magnitude is derived from total magnitude such as $AR = \frac{R}{Q}$, while the corresponding marginal magnitude is $\frac{\Delta R}{\Delta Q}$. According to calculus, if the units of R are continuous, and the change of Q (ΔQ) is infinitely small, it can be rewritten as $\frac{dR}{dQ}$ which is a derivative.

In geometry, the average curve is also derived from the total curve. The magnitude of any point on an average curve is the slope of a ray from the origin to the corresponding point on the total curve. Take point N in Fig7-1 as an example. The slope of ON is the ratio of the opposite side to the adjacent side, which is dividing Y-coordinate of point N by X-coordinate of it. Y-coordinate represents total revenue (R), while X-coordinate represents quantity (Q). So the diagram implies the same as $AR = \frac{R}{Q}$ in algebra. How about marginal magnitude? According to calculus, the derivative is the same as the slope of a curve in geometry. Therefore, the magnitude of any point on the marginal curve is the slope of the total curve.

After the introduction about the derivation of average and marginal magnitudes from total magnitudes in terms of algebra and geometry, let us turn to the relations among these three magnitudes. There are a series of propositions.

Let us continue to take the previous example of people's weight in a room. When a person walks in, his weight as a marginal magnitude is positive, and the total weight of the people in the room will rise. So the first proposition about the relationship between total magnitude and marginal magnitude is: When a total magnitude is rising, the corresponding marginal magnitude is positive. On the contrary, if a person walks out, his weight as a marginal magnitude is negative, and the total weight of the people in the room will fall. So the second proposition about the relationship

between total magnitude and marginal magnitude is: When a total magnitude is falling, the corresponding marginal magnitude is negative.

Then look at the upper diagram Fig7-1, where is the maximum value of the total revenue (curve R)? It is on the “top of the mountain”. What is the slope there? The answer is zero. That is why in the lower diagram in Fig7-1, the corresponding point in the MR curve is on the X-axis, which means the X-coordinate is zero there. This total curve is a parabola going downwards, but if it is a parabola going upwards, the minimum value of it is at the “bottom of the valley”, where the slope is also zero. So extreme value, no matter it is a maximum or a minimum, appears where the slope is zero on the total curve. In other words, when a total magnitude reaches a maximum or a minimum, the corresponding marginal magnitude is zero, which is the third proposition.

The previous three propositions are all about the relations between total and marginal magnitudes, and how about average and marginal magnitudes? Suppose a person walks in, but he now needs to be classified as “fat” or “thin”. If he is heavier than the average weight of the people in the room before he walks in, he is fat. On the contrary, if he is lighter than the average weight, he is thin. It is very reasonable. If a “fat” walks in (geometrically the marginal curve lies above the average curve), the average weight will rise. If a “thin” walks in (geometrically the marginal curve lies below the average curve), the average weight will fall. If a person who is neither fat nor thin (i.e. he is as heavy as the average weight) walks in, the average weight will not change.

To sum up, when the average magnitude is falling, the marginal magnitude must lie below it, which is the fourth proposition. On the contrary, when the average magnitude is rising, the marginal magnitude must lie above it, which is the fifth proposition. Finally, when an average magnitude is neither rising nor falling (at a minimum or maximum), the marginal magnitude equals the average magnitude, which is the sixth proposition.

Geometrically, the marginal curve passes through the highest or lowest point of the average curve. Fig.7-2 shows the relations among these three magnitudes taking total cost, average cost and marginal cost as the example. Let us focus on the lower diagram of average and marginal cost curves which are both parabolas going upwards. To the left of point L , the MC (marginal cost) curve lies below the AC (average cost) curve, which means $MC < AC$, so AC is pulled down by MC. To the right of point L , the MC curve lies above the AC curve, which means $MC > AC$, so AC is pulled upward by MC. The AC curve is falling to the left of L and rising to the right of L , so L is a turning point where is the “bottom” or the lowest point of the AC curve, and through which the MC curve passes.

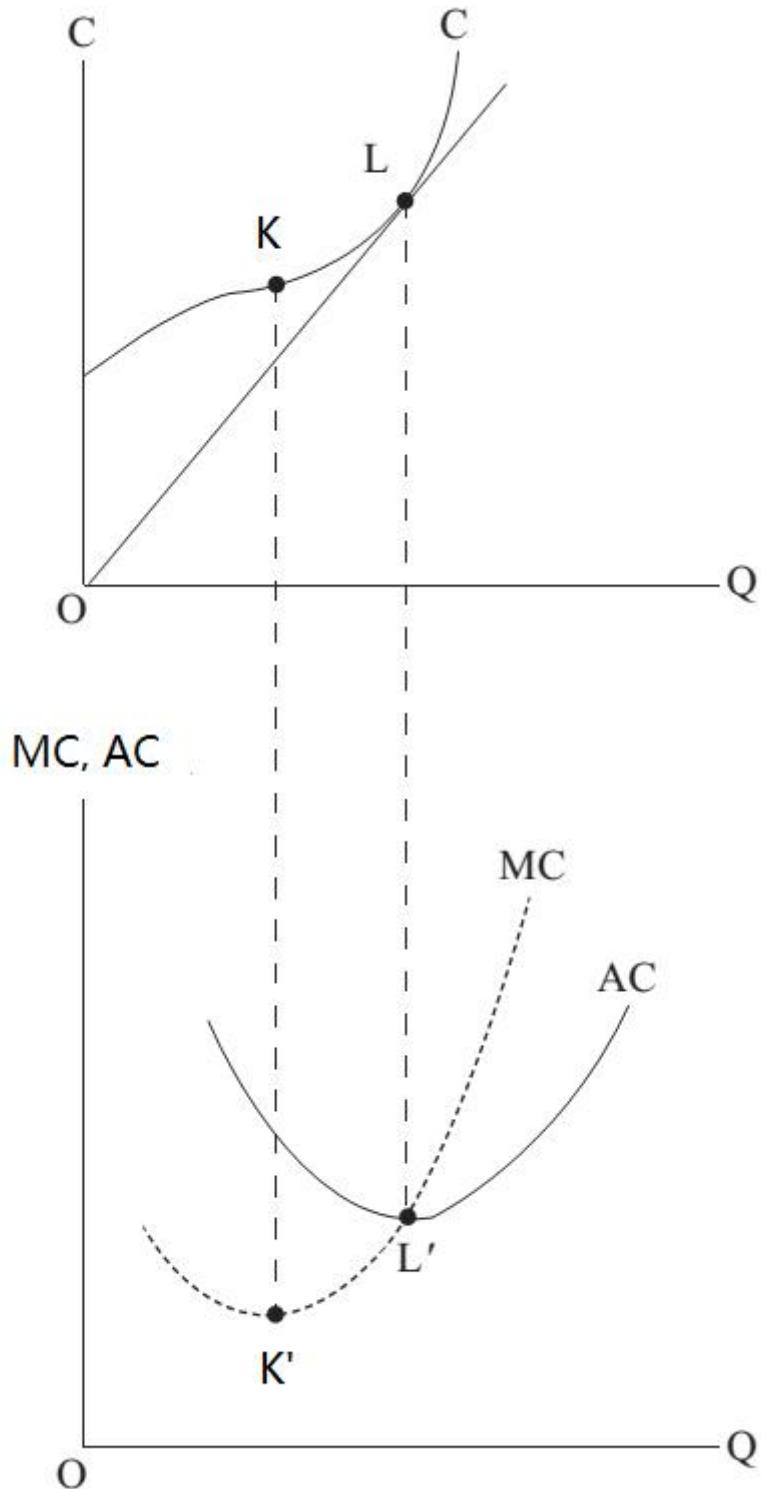


Figure 7.2

Readers can deduce by themselves the reverse case where both the average and marginal curves are parabolas going downwards. In the beginning, the marginal curve lies above the average curve, pulling it upward. Later, the marginal curve lies below the average curve, pulling it downward. The turning point, where the average curve changes from rising to falling, is the "top" or the highest point of the average curve, and through which the marginal curve passes.

Section 3: The Economic Implication of Marginal Analysis

In the previous section, the concept of margin has been introduced from the perspective of mathematics. However, margin is an economic instead of mathematical concept, or else it should not be called "margin" but "derivative". Margin is still significantly different from derivative in that there is requirement of "continuously differentiability" for derivative. As the example of the weight of people in a room in the previous section, that a person walks in is an incremental change of a discrete instead of a continuous variable because a person cannot be split into two halves, let alone into infinite numbers of infinitesimal persons. Strictly speaking, a discrete variable can not be taken the derivative. However, that is the requirement of mathematics, not economics which is not so demanding. As long as there is incremental change, no matter it is discrete or continuous, it is a margin.

As mentioned in Lecture 3, the postulate of self-interest cannot be directly applied to explain human's behaviors, but should be along with different constraints. If the constraints do not change, neither will human change his behavior, nor will the postulate of self-interest be meaningful. Economics is applied to predict that human will change his behavior according to both the postulate of self-interest and the change of constraints. By investigating whether human really change his behavior as predicted, economics including the postulate of self-interest is tested.

Self-interest is a postulate about the motivation of human's behaviors, and is unobservable intention. However, the change of constraints and human's behaviors are both observable facts. The postulate of self-interest and the change of constraints constitute the economic explanations which are tested by the change of human's behaviors. Therefore, if there is no change, there will be neither explanations nor tests. And change is margin! Difference is the same as change, which is also margin: that the same person changes his behaviors along with the change of constraints is logically similar to that different persons have different behaviors under different constraints.

Review the concept of margin from the perspective of mathematics introduced in previous section: MR or MC is the change of total revenue or cost due to the rise of Q. Similarly, from the perspective of economics, the change of human's behaviors is due to the change of constraints. This is the economic implication of marginal analysis, which can include all kinds of human's behaviors and constraints, and that is why marginal analysis is so important in economics.

Next, we will introduce various applications of marginal analysis in detail.

Firstly, the basic application of marginal analysis is to explain phenomena or behaviors by using the postulate of self-interest without falling into the trap of tautology.

For example, there is a question as homework I designed: a girl claims that she often offers her seat to others on the bus, so she is not self-interested or rational. One of the replies is, "The benefit of psychological comfort from offering her seat on the bus outweighs the loss from the tiredness of standing, so she chooses to offer her seat. By contrast, those who do not offer their seats think the benefit of psychological comfort is too little for them to outweigh the loss of standing." Is the reply right? How should you refute the girl?

The answer to this question is: the reply seems reasonable, but in fact it directly uses the postulate of self-interest to explain the girl's behavior, which is a tautology. In order to avoid making the mistake of tautology, the change of constraint must be added, which means marginal analysis must be applied. But how? The girl says that she offers her seat "often" instead of

"always". Why does she sometimes offer her seat but sometimes not? Is she sometimes self-interested or rational and sometimes not? It is obviously ridiculous. This is the best refute to the girl.

Of course, it is not the purpose of economics to refute the girl, but to explain her behavior scientifically. Sometimes she offers her seat and sometimes not, not because she is sometimes self-interested and sometimes not, but because the constraints she faces at different times are different, which leads to different behaviors. The girl will offer her seat on the short-distance bus, but how about on the long-distance train? Once she has offered her seat, she will have to stand for hours or even a whole day. Will she still choose to offer her seat? The girl will offer her seat to the old, children and pregnant women, but how about those who are strong and young? Will she still choose to offer her seat to them?

The above cases are not chosen randomly, but purposefully. The former case is about the change in the constraint of increasing cost, while the latter case is about the change in the constraint of decreasing income (the non-monetary income of psychological comfort). It is clear that with the change of constraints, the girl's behaviors have also changed, and the direction of the change is in line with the postulate self-interest, which means when the cost of offering her seat increases or the income decreases, she tends not to do so. Thus, her own behaviors have definitely refuted her claim that she is not self-interested or rational.

In a word, it is the basic application of marginal analysis: to add the change of constraint to the postulate, concepts or theories with the nature of tautology, so that the explanations using them become refutable and scientific.

Secondly, there are three extended application of marginal analysis.

(1) Marginal analysis can be used to screen the key constraints.

There are thousands of constraints in reality, but it is impossible for all of them to have critical impacts on the phenomena or behaviors to be explained. One of the important applications of marginal analysis is to help to screen the false constraints that seem to have impacts.

Specifically, observe whether there are different behaviors under the same constraint, or whether there is the same behavior under different constraints. If the constraint remains unchanged, neither should the relevant behavior change. On the contrary, the behavior will change along with the change of constraint. Therefore, if there are different behaviors under the same constraint, or if there is the same behavior under different constraints, that the relevant constraints may be redundant and not the key ones.

For example, since 2012, China has implemented the policy of free passage of cars with less than seven seats on expressways during the National Day. As a result, there has been serious traffic jam on expressways during the National Day. Different parties have explained this phenomenon differently. Economics is a science, and should avoid the disturbance of subjective feelings and interests to judge these explanations from an objective perspective. It is precisely where marginal analysis is very useful.

Some argue that the traffic jam on expressways is not due to the policy of free passage, but because there are always more cars using expressways during the National Day holiday than usual. In other words, in this explanation, the National Day holiday is regarded as the key constraint which causes traffic jam. However, there was the National Day holiday not only since 2012, which means this constraint has not changed at all. How can it cause the change in the phenomenon that traffic jam was significantly more serious than that in previous years? Thus, it has refuted this explanation of the National Day holiday as the key constraint.

Some others argue that it may be because the Mid-Autumn Festival and National Day in 2012 were linked together, resulting in more traffic than previous years. In this explanation, the long holidays are still regarded as the key constraint, but just the 8-day long holiday of 2012 is longer than before (7 days). However, with marginal analysis, this explanation is again easily refuted. If traffic jam was really caused by the longest holiday in history, this constraint is exactly the same for expressway and national highway. According to marginal analysis, if the constraint is the same, the phenomenon or behavior should be the same, which here means there should be the same traffic jams on national highways as on expressways. However, there was not such news of serious traffic jams in national highways.

(2) In economics, it is marginal magnitude instead of total magnitude that is important for explanation, so marginal analysis is more useful.

Different from natural science, it is almost impossible to obtain the total magnitude of a social phenomenon in social science. Fortunately, with marginal analysis, it is marginal magnitude instead of total magnitude that is important for explanation.

Let us continue to take the free policy on expressways during the National Day since 2010 as an example. According to marginal analysis, we can infer that the free policy is more likely to cause those who originally planned to go out to change from using national highways to expressways than to cause those who originally did not plan to go out to change to going out. When one changes from not going out to going out, the increase in marginal cost will not only include the toll of expressways, but also the expense on oil, food, housing and so on. However, if one has already planned to go out, the increase in marginal cost is only the difference of toll between national highways and expressways, which is actually negative with the free policy. Therefore, the free policy will cause all who have planned to go out to change from using national highways to using expressways, but just cause some who originally did not plan to go out to change to going out. We do not need to investigate how much precisely the total cost of going out is, but just need to logically identify the change in marginal cost.

(3) Equal marginal principle is a useful application of marginal analysis.

Equal marginal principle which will be discussed in the later lecture of "Market Structure: Price-Searching" is a very useful application of marginal analysis. It can be proved by mathematics that in order to achieve optimum, all relevant factors should be equal on the margin. For example, for producers, in order to achieve so-called profit maximization (strictly speaking, rent maximization), he should try to equalize the marginal revenue and marginal cost of the production (i.e. $MR=MC$), which is also called production equilibrium.

It is a more technical application of marginal analysis, and is widely used in mathematical models. However, even without mathematics, it can be also applied in economics. It can be inferred from equal marginal principle that it is optimal if this principle is satisfied, while it is not optimal if this principle is not satisfied.

Let us continue to take the free policy on expressways as an example. This free policy seems to be beneficial to people. However, the cost of expressways will not disappear, so the question of how to collect enough money to cover the cost remains. If there is toll for passage of expressways, the cost is covered by those who use expressways. By contrast, if it is free to use expressways, governments need to input fiscal money to cover the cost through tax collection. In the former case, those who use expressways pay the cost, so it is clear how much the marginal cost and marginal revenue are and if they are equal. By contrast, in the latter case, those who use expressways do not pay, but all tax-payers, no matter whether they have used expressways or not, pay the cost by paying taxes. It is not clear how much the marginal cost and marginal revenue are and if they are equal, because the users and payers are not the same people.

It is normally thought that the former case is fair while the latter case is unfair. However, economics forbids value judgment. Seen from economics, in the former case, it is easier to achieve the optimum that marginal income is equal to marginal cost, which means the transaction cost satisfying equal marginal principle is lower. By contrast, in the latter case, it is difficult to achieve this optimum, which means the transaction cost satisfying equal marginal principle is higher. Because the users and payers are not the same, the relationship between cost and income is not clear, which causes the so-called "waste" and is harmful to the interest of the whole society by using more social wealth or income to pay higher transaction cost.

It is also why there is higher transaction cost for indirect payment than direct payment. It is difficult for a consumer to know whether his consumption has reached the optimum where the marginal revenue is equal to the marginal cost, because the cost paid by him is not directly related to the revenue from his consumption. In other words, the information cost (a kind of transaction cost) about the optimal consumption is higher. In particular, consumers misunderstand that there is no cost when there is free policy, but in fact they have paid the cost indirectly by paying taxes, so they will consume too much, which means they consume more than the optimum determined by equal marginal principle, and the part of actual consumption exceeding the optimum is a waste with the cost greater than revenue.

Caution: even when there is no free policy implemented by government, the market will sometimes voluntarily chooses to use indirect payment, because under some constraint, it is possible that although the transaction cost of indirect payment is relatively higher, there is some other benefit which can cover the transaction cost, and indirect payment is optimal as a whole.

As for the example of free policy on expressways, those without cars are often poorer, while those with cars are often richer. It seems extremely unfair to pay the cost of expressways with taxes instead of collecting tolls, because it is equivalent to force the poor who have no cars and do not use expressways but have to pay taxes for expressways to subsidize the rich. However, the key point is those that infringe PPRs are evil policies or laws, no matter whether they infringe the PPRs of the rich or the poor. Therefore, the free policy on expressways is an evil policy not because it "robs" the poor to help the rich, but because it "robs" those who do not use expressways no matter they are poor or rich. No economic theory agrees with "rob the rich to give the poor".

Lecture 8: The Law of Demand

The postulate of self-interest has been introduced in Lecture 3. Just like geometry, axiom is derived from postulate. There is only one axiom in economics: the Law of Demand.

Section 1: The Law of Demand and its Key Points

The law of demand says, "Ceteris paribus (other factors remaining unchanged), less is demanded at a higher price, and vice versa." In terms of geometry, it is a curve sloping downwards. (See Fig8-1)

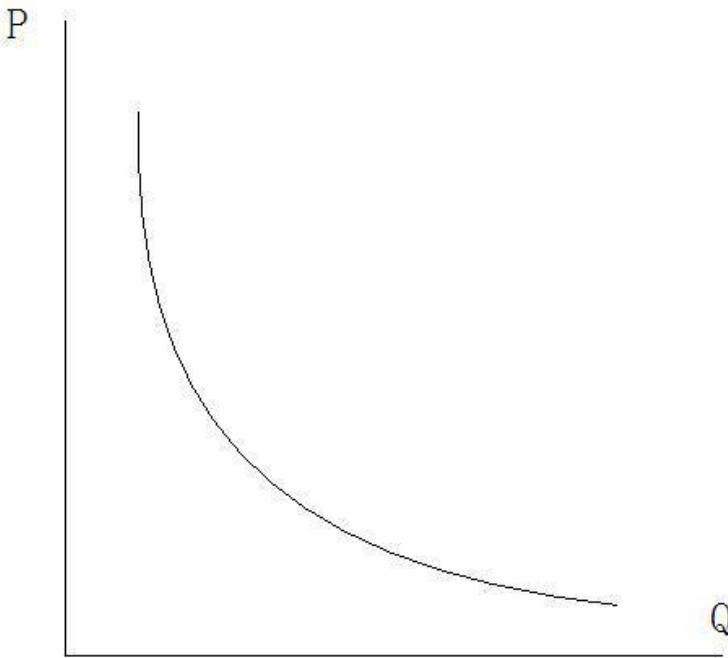


Figure 8-1

With respect to the demand curve, there are several key points need to be concerned.

(1) It is the change in price that causes the change in quantity demanded, not the other way around. In other words, price is an independent variable, while quantity demanded is a dependent variable. However, as can be seen in fig8-1, price is plotted on the y-axis and quantity demanded on the x-axis, which is exactly contrary to the common situation in mathematics where independent variable is plotted on x-axis and dependent variables is plotted on the y-axis. Why? There is no special reason, just because Marshall, who invented the demand curve, plotted price on the y-axis and quantity demanded on the x-axis, which is followed by all economists, thus becomes a convention.

(2) Price is an observable fact, but strictly speaking, it is not the nominal price or monetary price, but the relative price or real price expressed in other goods. For example, suppose there are two goods: a cell phone and a bottle of water. The cell phone is 1000 RMB and the bottle of water is 1 RMB, which are all monetary prices (nominal prices). The real price of the cell phone should be 1000 bottles of water, and that of one bottle of water is 1/1000 cell phone. As mentioned in Lecture 6, in price theory (or so-called microeconomics) there is only barter without money, so the price of one good is the relative price expressed by another good. Or although there is money, the value of money is stable, so the monetary price is not out of line with the real price. However, if the value of money is unstable, the monetary price will be out of line with the real price. For example, if the monetary price of a cell phone rises to 2000 RMB and that of a bottle of water to 2 RMB, their relative price is still 1000 bottles of water for one cell phone or 1/1000 cell phone for one bottle of water. In other words, the real price has not changed, only the money devalues

which means there is inflation. Price theory does not study the influence of money, so the price in it refers to relative prices (real prices).

(3) Quantity demanded is not an observable fact, but individual's intention, and is different from quantity of purchase which is a fact. Strictly speaking, the law of demand says, "Ceteris paribus, people tend to buy less when price rise, and vice versa. Whether this (psychological) tendency leads to real purchase depends on some specific constraints. Moreover, if quantity demanded is also a fact, the two variables of the demand curve are both facts. Applying the law of demand to explain phenomena is using a fact to explain another fact, which is not allowed by the methodology of science.

Section 2: Ceteris Paribus

There is "ceteris paribus" at the very beginning of the law of demand. Why? Because there are in fact many factors that affect quantity demanded, not just price. For example, preference and income are also factors that can affect quantity demanded. If one prefers to apple, he will tend to buy apples even at higher price. Or even if the price of apples remains unchanged, with higher income, one will usually tend to buy more apples. However, the law of demand focuses on the relationship between price and quantity demanded, so it is necessary to assume that other factors other than price, which may also cause changes in quantity demanded, remain unchanged, in order to separate the influence of price on quantity demanded. In some economic literatures, the variables studied by the theory or model are called endogenous variables, while the other factors that are not studied but also affect the dependent variables are called exogenous variables.

"Ceteris paribus" is very important, so it needs to be discussed in detail. Firstly, the law of demand assumes "Ceteris paribus", which is not like the postulate of selfish-interest which does not need to be proved, not allowed to be questioned, and will never change. Assumption is different from postulate, and is a prerequisite of the application of a theory. If the prerequisite of a theory has not been satisfied, even the prediction of the theory is inconsistent with the fact, it is wrong to claim that the theory is refuted by the fact. It is not the fault of the theory, but of those who wrongly apply the theory.

For example, suppose there is a vendor sitting by the street and selling umbrellas for 10 RMB. One day it suddenly rained, and he raised the price to 20 RMB. However, more umbrellas were sold out on that day. Does it refute the law of demand? Of course not, it is the weather as other factor that has changed. There is a series of specific other factors, including the sunny day, behind the demand curve corresponding to the price of 10 RMB. If it is rainy day, this other factor has changed, and the original demand curve of sunny day will no longer apply. Another new demand curve should take the place.

Suppose that D_1 in Fig. 8-2 is the demand curve of umbrella in sunny days and D_2 in rainy days. Behind D_2 there is a series of specific other factors which are the same as the initial D_1 except the "weather" changing from sunny days to rainy days. This new demand curve is obtained by moving the initial one parallel to the right, because at each price level, the quantity demanded corresponding to the D_2 is higher than that corresponding to the D_1 . For example, when the price is P , the quantity demanded corresponding to the D_1 is Q_1 and that corresponding to the D_2 is Q_2 , which shows that at every price level, due to the influence of rain, the quantity demanded for umbrellas is higher than in sunny days.

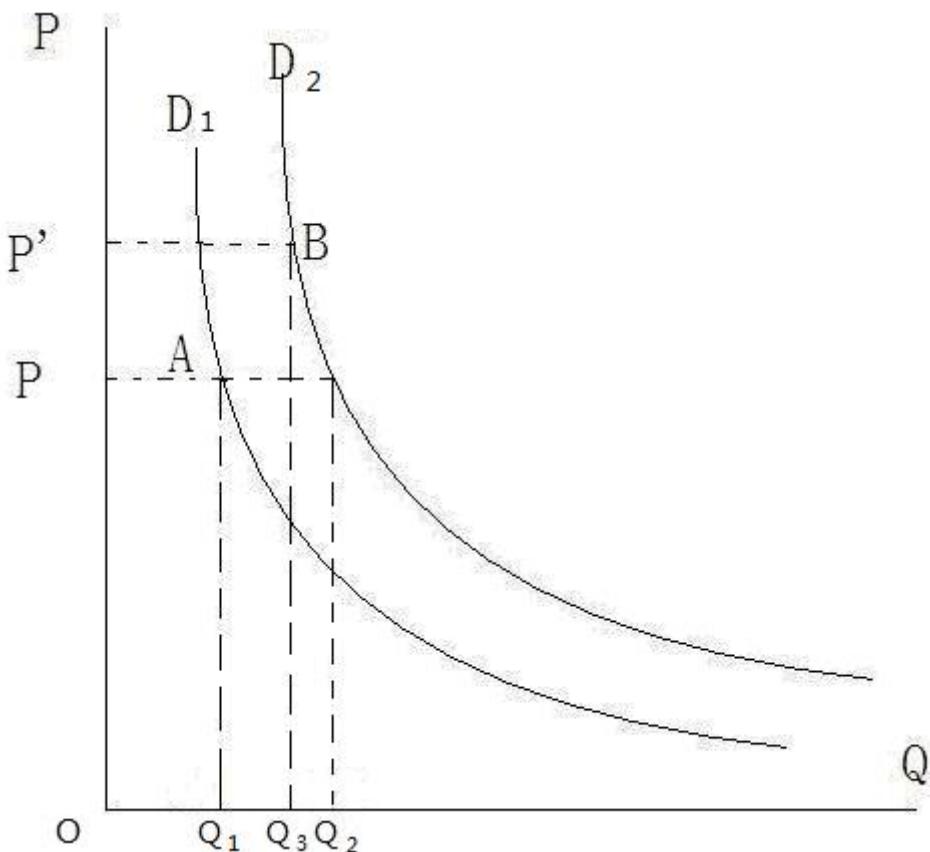


Figure 8-2

Those who neglect the prerequisite of the law of demand are wrong in directly connecting the point A and point B as a demand curve which slopes upwards. There must be prerequisites for any theory, which is also a part of the theory itself. Take an example in mathematics: when we say $1+1=2$, there is an implicit prerequisite of decimal system, but if binary system is actually applied, such as for computers, this calculation is wrong.

Therefore, the difference between "quantity demanded" and "demand" should be paid attention to. Geometrically, quantity demanded refers to the point on a specific demand curve, so the change in it means the point moving along the same demand curve. And demand refers to the whole demand curve, so the change in demand means the whole curve moving or the point jumping from one demand curve to another.

Many may think: hey, isn't it obvious? How can anyone make such a stupid mistake? Is it necessary to expound in such detail? However in reality, there are so many economists who make such mistakes, because the changes in other factors are sometimes not so easy as that of changing from sunny days to rainy days in previous example of umbrellas.

There was a big debate in economics of China about whether the demand curve could slope upwards which was dubbed as "the demand curve getting its tail up". Those who advocated that the demand curve could slope upwards are not only ignorant to economics but also to the methodology of science. The law of demand is the axiom of economics, which is the same as the postulate in that it does not need to be proved and must be accepted. If one does not accept the axiom, please get out and establish a new "economics with demand curve sloping upwards". What's more, it is not allowed to advocate that the demand curve sometimes slopes downwards and sometimes upwards, because it will make the law of demand become irrefutable tautology which seems to be able to explain everything but in fact explain nothing at all. When the fact is in favor of supporting the demand sloping downwards, it is convenient to say that now the demand curve sloping downwards apply, and vice versa. As a result, it is always true and impossible to be refuted by facts, but just becomes unscientific.

However, those who advocated that the demand curve could slope upwards argued that they had observed phenomena refuting the law of demand law or the demand curve sloping upwards. Really? Let's take three typical examples given by them, and judge by the law of demand.

The first example came from a post on an economics forum of a famous university when the debate was hot. The poster said that he saw with his own eyes a clothing store where a skirt was priced 20 RMB but could not be sold out, while later, when the price was increased to 200RMB, it was quickly sold out. Did it refute the law of demand? Of course not. But why? Because some other factors had changed. But what were they? It may not be so obvious in reality.

The answer is that other factors here are information costs. The buyer of the skirt certainly did not know that it had been priced 20RMB. If he knew and did not buy when it was priced 20RMB but chose to buy when it was priced 200RMB, his behaviors really refuted the law of demand, which exactly implies that the law of demand is refutable and scientific theory instead of tautology, because it can present the refutable condition. The poster must know that the same skirt had been priced 20RMB before, which meant his information cost was lower, so he would not buy it. He wrongly regarded the buyer's information cost the same as his, or wrongly regarded the buyer's demand curve the same as his, so was wrong in directly connecting the point on his demand curve and that on the buyer's demand curve to form a so-called demand curve sloping upwards.

This example needs to be further explained. With information costs which is one kind of transaction costs, consumers are generally not so good at judging the quality of goods in the market as producers, so they adopt the strategy of "judging quality by price" to reduce the relevant information costs. Generally the better the quality is, the higher the price is, which means quality is also a kind of other factors, there are different demand curves for better-quality goods and poorer-quality ones. Thus consumers tend to think that higher-priced goods are of better-quality. However, some sellers will make use of this strategy of "judging quality by price" of consumers to deceive them by deliberately setting a higher price for poor-quality goods. Of course, in the long run, market competition will eliminate this deception, because the information about the good will gradually spread which means the information cost will fall over time. It takes time for consumers to know the truth, which implies there are information costs. For a while, due to the constraint of information cost, some consumers were deceived into thinking that the skirt priced 200RMB was of better-quality than those priced 20RMB.

There are two reasons for the demand of the skirt increased after the rise in price. Firstly, the buyer did not know the skirt had been priced 20RMB before, and misjudged the quality of it by the price of 200RMB. Secondly, the demand curve for the skirts of better-quality priced 200RMB is located to the right of that for the skirts of poorer-quality priced 20RMB. In other words, the demand, instead of quantity demanded, for the skirts of better-quality priced 200RMB is greater than that for the skirts of poorer-quality priced 20RMB. As long as one of the above two conditions is not satisfied, the phenomenon seen by the poster will not occur. Take the poster himself for example: he knew the skirt had been priced 20RMB, so he would never misjudge the quality of it. Or if the skirt was priced 20,000RMB, consumers would doubt if a skirt of such top quality would be sold in a small store, so would not be so easily deceived. Even if some consumers were deceived to buy, the demand for the skirts of such "top-quality" would generally be lower than that of the skirts of poorer-quality priced 20RMB.

Next is the second example which came from an economist: the tuition of MBA (Master of Business Administration) was increasing year by year, but more and more applied for it. Did it refute the law of demand? Again of course not, and again because some other factors had changed which caused the demand curve of MBA to move to the right. It is logically similar to the above example of umbrellas, so Fig. 8-2 can be directly used to explain it.

What other factors had changed in this example? It is not so simple as the example of skirts just now. Only when you have a deeper understanding of the reality can you find out the key. Firstly, MBA was new to many at the very beginning, and few knew it well and did not know whether it was worth its tuition much higher than that of ordinary master degrees in business, so most would just wait and see, and few would try to apply. It is also due to the constraint information costs. However, people gradually got to know MBA over time, thus increased the demand for it, which meant the fall in information costs had moved the demand curve for MBA to right.

Secondly, the network of interpersonal connections is very important in doing business. Relatives, friends, fellows from the same town or province, classmates ... all can shorten the distance between people and are helpful for business. The majority of MBA students are businessmen, so it is certainly a good opportunity for them to form a network of classmates. In other words, MBA is to some extent similar to golfing. The value of education (or sports) in fact is not the most important, while it is much more important that it is a social platform for specific circle of people to form a network. With more and more students applying MBA, the value or the quality of MBA as a social platform was increasing, so its demand curve had moved to right.

To sum up, economics is usually not very difficult, but the application is rather difficult because we must have an in-depth and detailed understanding of the reality. Otherwise, only the prerequisite of "ceteris paribus" is enough to confuse many economic students and even economists. That's because they do not have a good grasp of the law of demand, but also because they do not have enough knowledge of the reality.

Section 3: The Explanation of “Buying on the upswing, Selling on the downswing” in the Stock Market

Last is the third example. There is a jargon called “buying on the upswing, selling on the downswing” in the stock market, which means investors tend to buy in when the stock price is up, while tend to sell out when the stock price is down. It seems that the more expensive, the more selling, and vice versa. Does it refute the law of demand? Again of course not. But why? Those who choose to buy a stock must expect that the stock price will rise, so the current price is lower than the expected future price, and vice versa. There are even no other factors that have changed in this phenomenon which just directly proves the law of demand.

What really makes an investor decide whether to buy in or sell out is not whether the current price is high or low, but whether the current price is higher or lower than the expected future price. Investors adopt the strategy of “buying on the upswing, selling on the downswing” not because they tend to buy the more expensive stocks and sell the cheaper ones, but the rising price is usually easier to make them expect the price will continue to rise and decide to buy, and vice versa.

However, if the rising price of a stock will only lead to the expectation that it will continue to rise, will it rise endless? It is vice versa for the falling price of a stock. Obviously there are not such things in reality? In reality, the rising price of a stock does not necessarily lead to the expectation that it will continue to rise. When the price rises to a certain level, some will begin to doubt if it has reached the peak, and reverse to sell out to take profits. The higher the price, the more expectations will be reversed. When the number of investors who reverse to sell out exceeds that of those who expect the price will continue to rise, the market will suddenly reverse. It is vice versa for the falling price of a stock.

The problem in explaining of “buying on the upswing, selling on the downswing” is that the constraint of rising or falling price of a stock is not enough to determine investors’ behaviors. In other words, with this constraint alone, we cannot reliably predict what investors will behave (decide). The key factor to determine the behaviors in the stock market is actually expectation. However, expectation is a kind of psychological activity, which is like self-interest and quantity demanded as motives and intentions, and cannot be directly used to explain phenomena. Psychological factors or activities are not observable, so cannot be tested by facts, which will make the explanation irrefutable and become tautology.

There are two solutions: one is to add observable constraints, which is the solution to deal with self-interest and quantity demanded (as will be explained later, price is the observable constraint); the other is to replace psychological factors with observable constraints, which is precisely the solution to deal with expectation. It is impossible to observe what one expects in his mind, but it is possible to observe the objective factors that form the expectations. For example, the implementation of a policy beneficial to a company will make investors expect the price of the company’s stock to rise. In principle, if one can know all the objective factors (constraints) that form investors’ expectations in the stock market, he can accurately predict whether the stock price will rise or fall, how much it will, and when it will. However, it is too difficult, in fact is impossible, to get to know all these constraints, which implies there is serious information cost (transaction cost). What is even more troublesome is that even if the information about the relevant constraints is available, they are changing every minute. Even one can know the relevant

constraints for this moment, they may have changed the next minute, so the formation about them soon become useless and new information needs to be collected again.

This is the characteristic of the stock market (in fact, all financial markets). There are not only many key constraints, but also they change rapidly which means after they change for once, they will change again in a short time. It is different from the usual decision-making in other fields where there are relatively less key constraints and they are more stable which means after they change for once, they will not change again for a long time. Is the stock market or stock price predictable? The answer is: yes, and no. Yes, because if one can quickly and accurately get to know all the key information, by applying the law of demand, it is enough to predict precisely. No, because in fact no one can get to know all the key information quickly and accurately. There is a school called "technical analysis" based on the charts of stock prices in the past (such as K lines) to guess the future trend. This so-called analysis actually assumes implicitly that "what happened in the past will continue to happen in the future" which has nothing to do with science. From the perspective of science, what happened in the past is based on some specific constraints, and what will happen in the future is based on other specific constraints. Why must these two sets of constraints be the same? No logic or facts can support this assumption. Remind: assumption is different from postulate in that it must be consistent with facts. Due to the too high information costs of getting to know the key constraints, some investors choose to guess based on the constraints in the past, which is understandable (explainable), but there is no reliable scientific basis for such guesses.

There are many complicated mathematical models in current finance. However, no matter how complicated the mathematics is, it cannot save the theories with wrong assumption. Mathematics can only ensure that the deduction from the assumption to the conclusion is logically correct, but cannot ensure that the assumption itself is correct. There is an idiom in China that if one go in the wrong direction, the faster he goes, the farther away he will be from his destination.

Strictly speaking, the financial market is difficult to predict, not because the financial theories are very complicated (as mentioned previously, the law of demand is sufficient), but because it is too difficult to get to know the key constraints. It is irrelevant to overcome this difficulty by applying very complicated mathematics. However, not only economics is helpless for the financial market, but physics is also helpless in some cases because the relevant constraints are too difficult to get to know. The most typical case is that physics cannot predict the outcome of gambling. Take the simplest coin toss as an example. As long as a physicist can accurately and quickly get to know all the relevant constraints such as the force of throwing, gravity, air friction, winds force, etc., he is sure to predict which side of the coin will face up by applying Newton's law. However, because the duration of the coin being thrown up and then falling down is too short, the physicist cannot get to know all the relevant constraints in time. All he can do is to calculate the probability of which side facing up in many tosses (50% for the positive and 50% for the negative), instead of predicting which side facing up at a specific time. In fact, all a financial economist can do is also to calculate the probability of a certain status, but the happening of small probability event can never be ruled out, and it is most horrible or even deadly which is called "black swan event" in the stock market.

From this comparison with physics, it is obvious that the short duration of constraints is the most critical difficulty. The constraints are short-term, which not only means that economists and physicists are unable to get to know all the key constraints in time, but also the information about the constraints quickly expire even if they manage to know. By contrast, insurance, which belongs to finance, also relies on probability, but not influenced by short duration. What's more, although there are also many complicated constraints in the investments in business, real estate, antique art, etc. other than in the financial market, neither is there the difficult of short duration of constraints. Therefore, economics as a science cannot predict the short-term fluctuations in financial markets, just as physics as a science cannot predict the results of gambling. Anyone who

tells you he is absolutely sure to make money in the stock market is a liar, because it is the same as anyone who tells you he is absolutely sure to make money in the casino is a liar.

Some may be still keep asking: if economics cannot predict the financial market, why are there so many securities analysts? And why are there so many people who spend money in enquiring them for forecasts about the stock market? Firstly, those who analyze based on the fundamentals of the companies and public news or inside stories actually investigate the long-term constraints. Of course, the key constraints that will influence the stock price cannot be only long-term ones, so such predictions based on the long-term constraints are helpless for short-term fluctuations. Secondly, as mentioned above, the school of so-called "technical analysis" does not investigate the relevant constraints, but only guess the future according to the past. In addition to the unscientific problem just pointed out, there is also the mistake of "explaining fact with fact" criticized by Marshall. However, this school in the financial market is not the only one who makes money by guessing. Fortune telling is the much older profession of the same nature. There have been always so many people who are willing to pay fortune tellers. What they buy may be just a peace of mind.

By the way, Burton Malkiel, a scholar of USA in his popular book "A random walk on Wallstreet" mocked that the return rate of the stock portfolio selected by a monkey with blinders randomly throwing darts would be almost the same as that of the stock portfolio recommended by securities analysts. In order to test this mock, a British magazine "Observe" held a stock-picking competition in 2012 among three teams: one is a cat, one is a group of securities analysts, and the other is a group of middle-school students. The result is that the cat did beat the other two groups of human by totally random selection.

Section 4: The Selection of Other Factors

-Methodology of Science (2): Occam's Razor

There are more complicated issues to discuss about "other factors." What on earth are the "other factors?" If "other factors" are any factors other than the price, then the question arises: if any factors other than the price remain unchanged, how can price change? What causes the price to change? So it is actually not simple for "other factors". Some other factors cannot change, but it is wrong that any factors other than the price cannot change. In other words, other factors should be divided into two categories: changeable and unchangeable ones. Next we will analyze which other factors are unchangeable, and which are as changeable as the price.

There are three kinds of other factors. The first kind of other factors is those that directly lead to changes in demand, and they are unchangeable. As mentioned in the previous examples, weather (which directly affects the demand for umbrellas), information, quality (which directly affects the demand for clothes, MBA, etc.) and income are all unchangeable.

The second kind of other factors is those that directly lead to changes in price, and they must be changeable, because if they do not change, neither will the price change. For example, when it comes to the end of season, the price of off-season clothes will be cut dramatically in clearance sale for funds returning. The qualities of the clothes have not changed, and the buyers also have sufficient information, so they will rush to do shopping. It is precisely the implication of the law of demand! Other factors such as seasonal promotion directly affect the price, so they must be allowed to change.

The third kind of other factors is called "intermediate factors", because they will be caused to change by the change in price, and then their changes will affect the demand. For example, the rise in the price of petrol will, on the one hand, make people buy less petrol (i.e. the demand for petrol drops), but it will also make people buy less cars because one does not need petrol if one does not have a car. Petrol and car are always used together, so the rise (or drop) in the price of one commodity will lead to the drop (or rise) in the demand for the other commodity. They are called "Complements". The characteristic of complements is that the price of one commodity is inversely related to the demand for the other. The drop in the demand for cars will lead to a drop in the price of cars, which can be explained by the TSD introduced in later lecture and can be understood even with common sense. As car is a supplement to petrol, the drop in the price of cars will in turn lead to a rise in the demand for petrol.

Another example: the rise in the price of pork will, on the one hand, make people buy less pork (i.e. the demand for pork will drop), but people still want to eat meat, so they eat chicken instead. Pork and chicken are substitutable in consumption, so a rise (or drop) in the price of one commodity will lead to a rise (or drop) in the demand for the other commodity. They are called "Substitutes". The characteristic of substitutes is that the price of one commodity is positively related to the demand for the other. The rise in the demand for chicken will lead to a rise in the price of chicken. As chicken is a substitute to pork, the rise in the price of chicken will in turn lead to a rise in the demand for pork.

As seen from the above two examples, the change in price of a commodity will lead to the change in demand and price of its supplements and substitutes which will in turn affect the demand of this commodity, so they are the "intermediate factors." Such intermediate factors must be allowed to change. Why? Because there are not only a large number of intermediate factors, but there are

also complex and subtle relationships with the goods analyzed. It is difficult to investigate whether they remain unchanged in reality. Take a real example in marketing. There was once a very successful chocolate advertisement in Britain, which attracted many middle school students. It was thought that only the sales of other candies and snacks would be affected because chocolate is obviously a substitute to them. Unexpectedly, even the sales of games dropped dramatically. After investigation, it was found that the pocket money of middle school students was basically a fixed amount, so when they bought more chocolate, they had to cut their spending on games as well as other candies and snacks. Game makers never thought that chocolate would become their competitor. In other words, any two seemingly unrelated commodities in this world will become substitutes or supplements under certain constraints. Therefore, if intermediate factors are not allowed to change, once one finds that the law of demand seems to be refuted by facts, one can always blame on the change of intermediate factors. However, it is difficult to investigate whether intermediate factors have actually changed, thus the law of demand will easily "degenerate" into a tautology without refutability.

From the above explanation on the three kinds of other factors, please pay attention to two key points. Firstly, that the first kind of other factors cannot change does not mean that they can really never change, but that they must remain unchanged on the same demand curve. If such other factors have changed in reality, we must draw another new demand curve. This key point has already been stressed before that "*ceteris paribus*" is an assumption which is not the same as the postulate of "self-interest". The postulate will never change and is always regarded as true unconditionally. However, the assumption is a prerequisite for the application of a theory, and it may not be true. It is true only when it is consistent with the facts, and then can the theory be applied.

Secondly, as mentioned before, there are other factors such as preference, income, etc. that affect the demand besides price. However, "preference" (as well as taste, like, and other similar expressions) should be specifically singled out. It does affect the demand, but one must assume that it will never change not only on the same demand curve, but also on different demand curves! Why? Because preference is a psychological factor, and not an objective fact that is observable. If it is allowed to change, again one can always blame on the change of preference whenever the law of demand seems to be refuted by facts. The psychological change is not observable by others, thus the law of demand will again "degenerate" into a tautology without refutability.

Take the previous example of umbrella, with the rise of the price, there are still more umbrellas sold out, which is explained by the change of objective fact (other factors) that the weather changes from a sunny day to a rainy one. However, it can also be explained by the change of psychological factor that umbrellas are preferred now than before. Although the effect of these two explanations is the same: to draw another new demand curve (both factors directly affect the demand and cause the whole demand curve move), the former explanation can be tested by looking up to see if it is really raining, but how can the latter explanation be tested? Who knows if umbrellas are more preferred now than before. In fact, it is logically similar to that economics forbids using the postulate of sometimes self-interest and sometimes un-self-interest to explain phenomena.

To extend the above analysis, the statement that a person's preference remains unchanged is logically equivalent to the statement that different persons have the same preference. In other words, it is forbidden in economics to use different preferences of different persons to explain their different behaviors. It is said that the Mongolian with much less income than Chinese consume much more than Chinese. It is easy to think that it is because the Mongolian and the Chinese have different consumerism, which is not a scientific explanation. Why the Mongolian has the consumerism different from the Chinese is itself a phenomenon that needs to be explained. In brief, such psychological factors as preference habit, custom, nationality and etc. cannot be

used to explain phenomena, but are phenomena to be explained. It is objective facts (constraints) behind these factors that can be used to make a refutable explanation.

It is somewhat similar to the previous discussions of expectation and intermediate factor. Intermediate factor is fact, but the change of it is too complicated to investigate, so it is better to let it change. Expectation and preference are the same in that they both belong to psychological factors which are not facts, and they should be replaced by the objective factors (constraints). In other words, the psychological factors are assumed to be unchanged, and it is the objective factors replacing them that change.

Take the previous example of umbrella again. One explanation is that because the weather changes from a sunny day to a rainy day, the whole demand curve for umbrellas moves. Another explanation is that because the umbrellas are preferred more in a rainy day than in a sunny day. However, preference is a psychological factor. In order to avoid the mistake of tautology, we should further look for objective factors (constraints) that cause this psychology. It is obvious that umbrellas can protect one better from adverse weather in a rainy day (one will get wet in rains without an umbrella, and has a greater chance of getting sick). Rain and illness are objective factors that can be tested by investigating the facts. Thus, the psychological factors that are not facts can be replaced by the logically related objective factors that are facts, and we can avoid the mistake of tautology.

Both of the above explanations are refutable and scientific. Obviously, the latter explanation is too long-winded. Here we introduce more requirements for methodology of science: a scientific theory should be not only refutable, but also the simpler the better. In science, there are also competitions among different theories, and there are two competition criteria to decide which one is the winner. The first criterion is "explanatory power" which means the ability to explain as many phenomena as possible without being refuted by facts). It is the theory with the strongest explanatory power that is the winner. If more than one theory can explain the same amount of phenomena, which means their explanatory powers are the same, now it is the second criterion of "simplicity" decide the winner. It is the simpler theory that is the winner. This second criterion in methodology of science is called "Ockham's Razor" because it was proposed by William of Occam (1287-1347), a logician and Franciscan friar from the 13th to 14th centuries. "Occam's Razor" says "do not multiply entities without necessity", while the common expression is "the simpler theory, the better."

It is necessary to discuss again the competition between Heliocentric and Geocentric in astronomy. In fact, Copernicus did not intend to reject but to improve Geocentric originally. With the development of astronomical observation technology, more and more phenomena that seem to refute Geocentric are discovered, which means Geocentric cannot explain those phenomena. When a theory is refuted by facts, scientists will not immediately abandon it and start to establish a new one, but try to amend and improve the existing theory. Astronomers at that time (including Copernicus) also did so by constantly adding new contents to Geocentric to explain the newly discovered astronomical phenomena. So Geocentric had not really been refuted at that time, but the problem is that it had become more and more complex which was not consistent with "Occam's Razor". If there are no other theories that can explain as many astronomical phenomena as it, astronomers can only endure the increasing complexity of Geocentric and continue to accept it. However, Copernicus observed that, unlike the distance between the earth and other stars would change, the distance between the earth and the sun almost never changes. Logically, there are only two possibilities: either the earth is stationary and the sun revolves around it, which is what Geocentric advocates, or the sun is stationary and the earth revolves around it. As the Geocentric became more and more complex, Copernicus changed his thinking: What if the sun is stationary and the earth revolves around it? Once this assumption was changed, he found that the explanation of the movements of various stars suddenly became extremely simple -- much simpler than Geocentric! Although all the movements of stars observed at that

time could be explained by both Geocentric and Heliocentric, the latter is much simpler, so "Occam's Razor" cut off Geocentric. In other words, strictly speaking, Geocentric was not refuted by facts in Copernicus era, but replaced by Heliocentric that is simpler with the same (or even stronger) explanatory power. It was not until Galileo's era that the phase of Venus was observed and Geocentric was definitely refuted.

Therefore, it is not a tautology to use preference to explain phenomena if the objective factors that cause preference is further found, but it is better to directly use those objective factors as explanation.

Through the above detailed discussion, it is hoped that the readers have not only learned the specific knowledge about the law of demand, but also further understood the "refutability" and "Occam's Razor" in methodology of science.

Section 5: The Role of the Law of Demand as An axiom

What is the relationship between other factors in the law of demand and the constraints mentioned before? In reality, human's behavior (choice) is constrained by many factors. Think of Rousseau's words: Man was born free, and everywhere he is in chains. If the so-called chains need a clear definition, they are constraints. There is such thing as unconstrained situation, so neither is there is such things as unconstrained freedom.

From the point of view of economics, self-interest is the most important restriction on human's behavior, but it is a postulate which needs not to be true. Only self-interest is not enough to constrain human's behavior. It has been pointed out in Lecture 4 that if there were no scarcity, human would not need to choose even if they were self-interested. There are also constraints that restrict human's behavior along with self-interest. Constraints are objective facts which are essentially different from self-interest. It is the unchanged postulate of self-interest and changes in constraints that cause human's behavior to change. However, there may be tens of thousands of constraints that cause human's behavior to change. It is impossible and unnecessary for economics to care about all of them. Only some of constraints need to be singled out to analyze which is called endogenous variables (such as price in the law of demand), while other constraints should be treated as exogenous variables (other factors).

It has been repeatedly stressed that it is the postulate of self-interest and changes in constraints that are used to explain phenomena in economics, which is precisely the law of demand. The price in the law of demand is a fact, so if widely interpreted, it can be regarded as constraint. Quantity demanded is individual's intention, so it is a psychological factor, not a fact. There is a negative correlation between price and quantity demanded, or the demand curve slopes downwards, which implies the postulate of self-interest. That is why in economics the demand curve must slope downwards and no exceptions are allowed (no demand curve can slope upwards), because if there were exceptions, it would be inconsistent with the postulate of self-interest.

Thus, the law of demand is not only a theory which can only explain the consumers' choices in the product market, but also an axiom of economics, and has strong explanatory power. Price can be regarded as constraint, so any constraint that plays a key role in determining the behavior to be explained can be put into the law of demand as price. When there is more than one key constraint, but just one of them is singled out to analyze, the other constraints will be regarded as "other factors" to keep unchanged. In fact, all the phenomena that have been explained previously with the postulate of self-interest and changes in constraints can be rewritten with the law of demand. The law of demand may not be mentioned literally, but it has actually been always applied and followed in any explanation or analysis in economics. And only such economic explanation or analysis is correct when it is consistent with the law of demand, the axiom of economics.

Lecture 9: Supply Curve, the Theory of Supply and Demand, the Price Intervention by Government

Section 1: The Supply Curve

The last lecture is about the law of demand, while this lecture turns to the supply curve.

Supply curve and demand curve are actually the same curve. Since the theory of producers has not been introduced, let's temporarily assume a world without production. In the absence of production, why are supply curve and demand curve the same curve? Suppose ABC in Fig. 9-1 is a demand curve for an individual. The point B means that when the price is O'B, the quantity of the good he holds is OO' (caution: O is the original point where the quantity demanded is zero). If the price drops from point B, according to the law of demand, he will increase the quantity demanded for the good, which means he moves along BA. On the contrary, if the price rises from point B, he will reduce the quantity demanded. But how? The answer is: sell the good he holds. That is supply! Using O'B as a mirror, the inverted line of BC can be drawn as BC' which is exactly a supply curve sloping upwards. So it is clear that supply curve is actually the inverted image of demand curve.

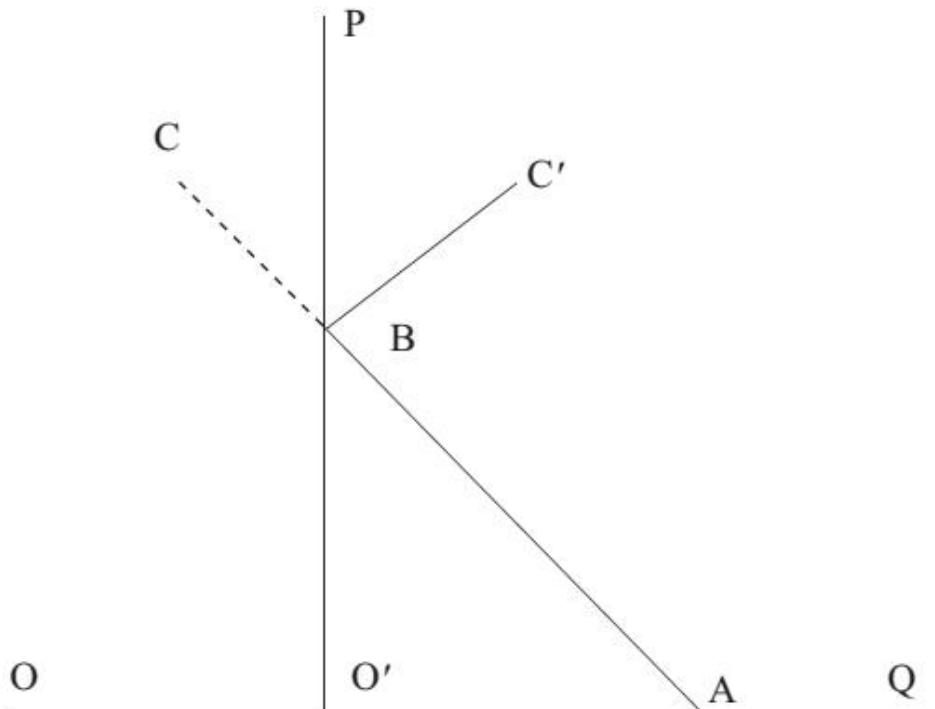


Figure 9-1

The economic implication of Fig. 9.1 is whether a person is a buyer or a seller, a consumer or a supplier depends on the price. If the price is low enough, people will choose to be buyers, and the lower the price is, the more they will buy. On the contrary, if the price is high enough, people will choose to be sellers, and the higher the price is, the more they will sell. Supply and demand can transfer from one to the other, depending on the price.

Since supply curve and demand curve are actually the same curve, the characteristics of supply curve are the same as those of demand curve, such as (1) quantity supplied is an intention, while the quantity of sale is a fact; (2) quantity supplied is different from supply, and the change of the former refers to the movement of a point on the same supply curve, while the change of latter refers to the movement of the whole supply curve. Readers should be able to analogize all of them from demand curve.

Section 2: The Theory of Supply and Demand

Combine the demand curve and supply curve, the theory of supply and demand is deduced. As shown in fig. 9-2 where the demand curve and supply curve of different people for a good are drawn in the same graph, the intersection point is P^* , the equilibrium price which determines Q^* , the equilibrium quantity. Caution: it is the price that determines the quantity, not the reverse, because the price is an independent variable and quantity is a dependent variable. Of course, it has stressed in Lecture 7 that there is problem for the concept of "equilibrium". Specifically, quantity demanded and supplied are both intentions. Are they equal to each other? God knows. My opinion is: it is always in equilibrium, so quantity demanded is always equal to quantity supplied.

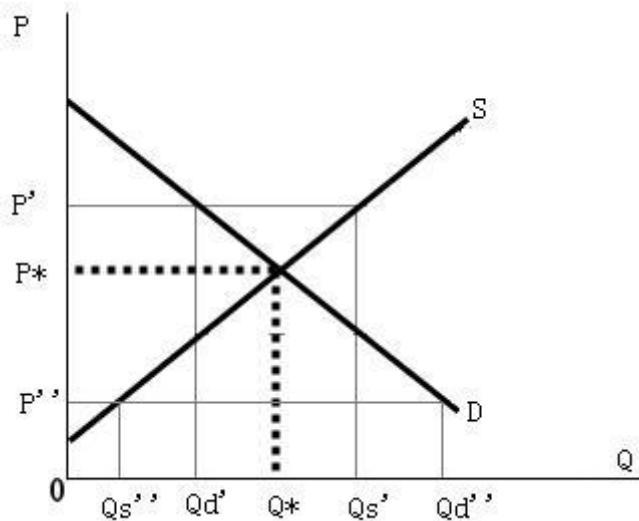


Figure 9-2

There is no such thing as disequilibrium, but in order to understand why P^* is the equilibrium price, just imagine the so-called disequilibrium when the price is not P^* .

There are two possibilities of disequilibrium, one is the case when the price is higher than P^* , the other is the case when the price is lower than P^* .

When the price, such as P' , is higher than P^* , the sellers are so willing to sell more of the goods that the quantity supplied is Qs' , while the buyers are so reluctant to buy that the quantity demanded is Qd' . There is oversupply in the market, and only the quantity of Qd' can be sold out, while the quantity of Qd' Qs' cannot. The diagram only tells us that the quantity that can be sold out at the price of P' is Qd' , but it does not tell whose goods are the "lucky ones" that are those sold out. In other words, the sellers must compete to become the lucky ones. Suppose the qualities of the goods are the same, the sellers can only compete with price by cutting prices to attract buyers, so the price of P' cannot be stabilized and fall. As the price falls, the sellers' desire to supply is hit, and some reduce supply, and some even leave the market completely, so the quantity supplied decrease. On the other hand, the buyers' desire to buy is stimulated, and some has bought increase demand, and some has not bought become to want to buy, so the quantity demanded increase. Thus, the oversupply in the market decreases. However, as long as there is still oversupply, there will be some goods that cannot be sold out, and the competition among the sellers will push them to continue to reduce the price. The price cannot be stabilized until it

decreases to P^* . Until then the quantity demanded and supplied are both Q^* , and all goods can be sold out, so the sellers do not need to reduce the price any more.

On the contrary, when the price, such as P'' , is lower than P^* , what will happen? Logically similar to the above analysis, it can be deduced that there is over-demand in the market, and only the need within the quantity of Q_s'' can be met. Now it is the buyers who have to compete to become the "lucky ones" by increasing the price. As the price rises, the buyers' desire to buy is hit, while the sellers' desire to supply is stimulated, so the over-demand in the market decreases. However, as long as there is still over-demand, there will be some buyers' needs that have not been met, and the competition among the buyers will push them to continue to increase the price. The price cannot be stabilized until it increases to P^* .

The above process from disequilibrium to equilibrium is just imagined for logical reasoning. In reality, it is always in equilibrium. If it seems to be in disequilibrium, it must be due to the neglect of transaction costs for both the buyer and seller in adjusting their behaviors, including changing the price or quantity. If there were no transaction costs, all adjustments would be completed in an instant. Precisely because there are transaction costs, it is in equilibrium in every instant.

When other factors have changed, the whole demand curve or supply curve will move, causing the equilibrium to change. In the previous lecture, a large number of such examples have been explained, so it will not be discussed here anymore.

Section 3: The Traditional Analysis of Price Control and Its Problem

There are two kinds of government intervention in price: price ceiling and price floor.

"Price ceiling" refers to the government's policy or law to repress the price to be lower than the equilibrium price. The price ceiling must be lower than the equilibrium price, otherwise it is meaningless. Price ceiling is also called "price control".

In contrast, "price floor" refers to the government's policy or law to inflate the price to be higher than the equilibrium price. Price floor must be higher than the equilibrium price, otherwise it is meaningless. Price floor is also called "minimum price" or "price support".

Both price ceiling and price floor are government interventions in price, which are essentially infringements of PPRs. As mentioned briefly before, there are three rights in PPRs: the use right, the right to income and the right to transfer. Among them, the use right and the right to transfer are actually needed to establish the right to income, because only when people can freely use and transfer the goods they hold, can they fully obtain the income from the goods. It is the right to income that is directly infringed by government intervention in price. This good belongs to me. I want to trade it at this price. On what legal basis can the government forbid me? Isn't it an infringement of my right to income?

Price control infringes the seller's right to income, while price floor attempts to infringe the buyer's right to income. To say "attempt" means it is actually not successful. Before the transaction, the PPRs of the goods were in the hands of the seller instead of the buyer. The buyer foresaw that his right to income would be infringed after the transaction, and he simply chose not to buy to avoid infringement. Therefore, price floor actually infringes the seller's PPRs, just not on the right to income right, but on the right to transfer because the seller cannot exercise his right to transfer at such a high price. Of course, when the right to transfer cannot be exercised, the right to income will be inevitably affected, because unless it is in an autarky world where there is no trade, the income from goods is mostly realized by transfer (sale).

Let's analyze price control firstly.

Fig.9-3 is the geometric analysis of the consequences of price control in MSE. The equilibrium price is P^* , but the government imposes price control to repress it to P' . At such a low price, sellers are reluctant to supply, so the quantity supplied is only Q_s' . However, buyers are very willing to buy, so the quantity demanded is as high as Q_d' . There is over-demanded in the market, which is called "shortage". As a result, only the need of Q_s' can be met. In other words, the consequence of price control is that fewer buyers' needs are met, while sellers' incomes drop, which can be described as loss at both sides.

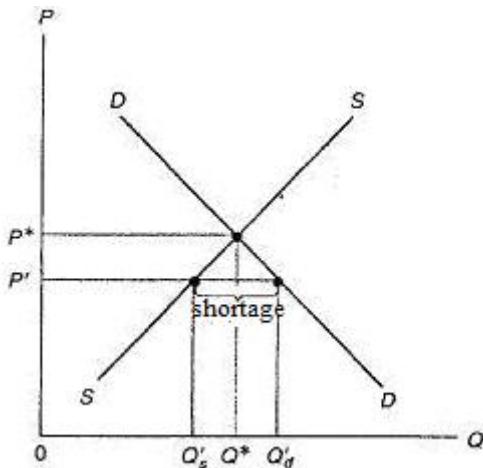


Figure 9-3

However, this analysis of MSE is wrong. Because quantity demanded and supplied are both intention, only God knows whether they are equal to each other, and there is no way to test which is more or less than the other. In other words, the so-called “shortage” is not a fact. Attention: the concept of “scarcity” mentioned in Lecture 4 is a fact, because the scarce goods have prices, and the scarcer a good is, the higher its price is, which are all observable facts. But what about “shortage”? There are no observable facts about it. What can be observed in reality is the quantity of purchase and sale, not the quantity demanded and supplied. However, the quantity of purchase is always equal to that of sale, no matter there is price control or not, because they are actually the quantity of transaction: from the buyer’s point of view it is the quantity of purchase, while from the seller’s point of view it is the quantity of sale. Therefore, the analysis on price control of MSE is logically wrong.

Next, the conclusion of analysis of MSE will be tested by facts. The conclusion is that price control will lead to a decrease price and in quantity demanded which should be referred to quantity of transaction. But really?

Suppose a landlord rents out a house, and receives RMB1000 a month. Now all of a sudden, the government imposes a rent control of RMB800 a month which is said to help the poor afford to rent a house. What will the landlord do? The economic implication of the analysis of MSE is that he will submit to the price control, but only rent out 80% of the house, which is obviously refuted by facts. In reality, the landlord will try to go round the price control, and get the original price back. There are many ways to achieve it, and the simplest way (with the least transaction cost) is: the landlord rents the house by RMB800 a month in name only, but charges RMB2400 of “contract fee” before signing a one-year lease contract with the tenant. Who refuses to the contract fee, the landlord will not rent the house to him, so he actually collects the same rent of RMB12,000 a year as before. The problem is if the tenant will accept the extra contract fee. If he accepted the rent of RMB1000 a month before, he would certainly accept the extra contract fee, because the contract fee shared over the 12 months of a year is RMB200, the sum of which and the nominal rent of RMB800 is no different from the original RMB1000. Moreover, what is more important is that in fact there cannot be houses worth RMB1000 in the market that are currently rented in RMB800 without any contract fee. As long as a landlord think of the idea of contract fee to go around the price control and get the rent back, others will get to know, so all landlords will do so, the tenant can only accept the extra contract fee, just as he accepts the equilibrium rent in the market before the imposition of price control.

In reality, there is really contract fee. There is rent control in New York. As a result, a tenant has to pay a sum of money called “key money” to buy the key to the house when he moved in, in addition to the namely low rent. How can a tenant enter the house without the key? Therefore, the “key money” is essentially the contract fee mentioned above. There had also been a case in New York that a tenant was required to buy a broken chair at a ridiculously high price which was

actually the contract fee again. The Landlords would use various names for contract fee by which took back the rent infringed by government control.

Similar phenomena occurred in Hong Kong when rent control was imposed. The landlords in Hong Kong were so imaginative that they invented "shoe money". It meant that the landlords had been walking around to find the suitable tenants and worn out their shoes, so the tenants should pay them to buy new shoes. Of course, the price was entirely not determined by the actual costs of the shoes themselves, but by the rent infringed by government control.

Let's look at another example in reality. There had been oil shocks in the 1970s, and the US government had imposed price control on petroleum for four times in an attempt to curb inflation by suppressing the rise in petroleum price. However, this was actually a drug rather than an antidote, because once the price control had been abolished, the petroleum price would rebound sharply, causing the consumer price index (CPI) to rise again. During the period of price control on petroleum, it was common for cars to line up to wait for refueling. And the petrol stations had also come up with a new trick: cars that refuel and lubricating can have priority to refueling. Only petrol was subject to price control, while lubricant was not, so the petrol station increased the price of lubricant so that the regulated part of the price of petroleum was transferred to lubricant. It is a kind of tie sale of a good with price control and another good without price control to transfer the regulated part of the price of the former to the latter.

Some may suggest government should also impose price control on lubricant. It can be expected that the petrol station will change tie sale of petroleum and car washing. What if government also imposes price control on car washing? Again, the petrol station will change to tie sale petroleum and the shopping in the store of the station. Unless the petrol station is nationalized to a state-owned enterprise, it can always "invent" some kinds of tie sale to get back the regulated part of price of petroleum.

It should be always remembered that human is self-interested and there is scarcity everywhere, which cannot be changed by price control, so neither can it be changed that there must be competition criterion to decide winners or losers. Without price control, price criterion is used to determine who can get those scarce goods in competition. With price control, part of the price criterion has been abolished, but the question of who can get the scarce goods by competition is still there. Under different constraints, people will use different ways to solve this question, sometimes going around price control and back to price criterion (such as contract fee or tie sale), sometimes using other non-price criteria (such as queuing) instead. There are numerous ways, which will lead to the "unconstrained" problem mentioned in the introduction of methodology of science in Lecture 2. Therefore, the key of economic analysis on price control is not shortage the MSE is concerned with, but the law that determine what ways people will use to take place of price criterion abolished by price control.

Under price control, part of the sellers' rights to income are infringed by government, but government just forbid them from collecting that part of the price (income) in the former way, and does not clearly stipulate who owns that part of income, so self-interested people (mostly the sellers, but sometimes the buyers) will try their best to grab it (the seller would think that he only took back the income he deserved). In fact, self-interested people are not competing for the scarce goods, but this part of income, because who is the owner of it becomes unclear due to price control. Of course, it must cost much to grab the income, which implies that there is transaction cost. It is also determined by the postulate of self-interest that the way that could grab the most income with the least transaction cost will be used. This is precisely the law that the theory of price control should seek to explain people's choice under price control.

In conclusion, price control is a very complicated issue, involving the ownerless income due to the partial infringement of PPRs, so self-interested people compete for grabbing it which leads to rent dissipation (a kind of transaction cost). However, self-interested people will also try to reduce the loss rent dissipation as much as possible. This issue will be explained once more in detail in the following lecture of "Rent Dissipation".

Section 4: Price Control Causes Transaction Cost to Increase

As discussed in the previous section, the analysis of MSE focuses on the so-called shortage is meaningless and has been refuted by facts. There are two other points should be paid attention to about price control. Firstly, some goods are not paid directly at visible price, but indirectly in other ways. If government controls those indirect ways, it is also price control. For example, the time that TV audiences spend on watching advertisements is an indirect price to pay for TV programs, so it is actually a kind of price control if government imposes restriction on TV advertisement.

Secondly, some think that government regulation is always invalid, which is true seen from the example of rent control. The rent actually paid by the tenant is still RMB1000 a month which is the same as the equilibrium price originally and just is not all in the name of rent, but part of RMB200 as "contract fee". Judging from whether the government regulation has achieved its original stated goal, price control is totally a failure, so it is an "evil policy" and should not be implemented.

It has been discussed in detail in Lecture 5 that the question of "ought to be" belongs to moral ethics, and should not be answered by economics. However, if an objective standard can be specified in advance, and based on which the conclusion of empirical analysis is judged, then it has become normative analysis in economics. Here is a general objective standard for all the normative economics to judge if a government policy is good or bad, that is, if the objective effect of a policy is the original goal it claims to achieve. If it is, it is a "good policy", while if it is not, it is an "evil policy". The rent control is a typical example: government claims to implement this policy in order to decrease the rent burden for the poor. However the empirical analysis as mentioned in previous section shows that the tenant pays no less than before. So a normative conclusion can be come to that rent control is an "evil policy" that should not be implemented.

In other words, to judge if a policy is good or bad does not depend on what good purpose it claims to achieve, but on what actual effect it results in, which is also a philosophical truth. Economics is a science, but what we can learn from it is not only economic theories, but also the way of thinking and even philosophical truth. Here is: do not judge a person by what he says, but what he does! Do not believe what a person claims as subjective motive of his behavior which is neither credible (only God knows if he has lied) nor important, but observe the objective effect of his behavior. You say you love me, but you love me so much that you kill me! I am afraid that I do not want such love.

Some may still argue that the effect of rent control is not good just because it is not perfect enough and government should improve the policy by forbidding all kinds of "contract fee". However, the example of price control on petroleum has shown that the essence of price control is that it infringes PPRs and makes part of the income ownerless. Self-interested human will always try all the means that government never dreams of to grab the ownerless income. The so-called improvement of policy cannot change the relevant key factors of self-interest, scarcity and competition.

And even if government "improves" the price control such that no one can grab the ownerless income, the result is only all the regulated rent dissipates. How can be it good to society? For example, if government has regulated in great detail and nullified all means for collecting "contract fee", which means in essence that the transaction cost of contract fee increases to be even higher than the regulated rent, what will be the result? Before rent control when the landlord could get RMB1000 a month from the house, he would furnish the house with furniture.

Now with rent control, he can only get RMB800 a month and cannot get back the other rents by "contract fee", so he will only rent an empty house without furniture. Such an empty house could only be rented out in RMB800 a month even when there was no rent control. Thus government's rent control is still invalid. Or the landlord will no longer repair and maintain the house from now on, so that the house will depreciate to the level consistent with the regulated rent. All the houses worth more than RMB800 a month in the market will eventually disappear, so those who are willing to pay more rent to improve their living conditions no longer find a house good enough to meet their needs even they have money. Does government really want to achieve such a result by rent control?

In addition, the government regulation is invalid does not mean that it has no adverse impact on society. From the previous analysis, it can be seen that if government strongly enforces rent control and effectively forbid all kinds of ways going around the regulation such as contract fee, there will be only very bad consequences to society. However, does government regulation have no negative impact on society when there is contract fee? Self-interested people will compete for the ownerless income due to price control, which will cause transaction cost. Self-interested people will try to minimize the transaction cost which is not zero, so there is an increase in transaction cost in the regulated society compared with the unregulated one.

Take rent control as an example. When there is no rent control, the cost for the house is RMB1000. When there is rent control, the cost for the house is the sum of RMB800 in the name of rent and RMB200 in the name of contract fee. It seems that there is no difference. However, when there is no regulation, there is only one price to be set by the landlord and tenant (i.e. rent of RMB1000). After rent control, there are two prices to be set (i. e. rent of RMB800 and the monthly contract fee of RMB200), so there is more transaction cost than before. Of course, the increase in transaction cost here is almost negligible, because the rent of RMB800 is set directly based on the government regulated price, while the contract fee of RMB200 is set based on the difference between the market rent and the regulated price. It is precisely why contract fee is chosen as the least costly way to go around the rent control.

Nevertheless, it is only because in the house leasing market it is easy to find this way that there is little increase in transaction cost under certain constraints. In many cases, it is not so "lucky" under other constraints. Let's turn to the example of price control for petroleum. The gas station get back the regulated price of petroleum by tie sale of it and lubricants and increasing the price of lubricants, which increase much more transaction cost than contract fee in the house leasing market. A car that does not need to be lubricated at all will have to be done so in order to refuel without queuing. It is obviously a waste to be more frequently lubricated than necessarily, which implies transaction cost. And some will choose not to let their cars be lubricated but to queue up. By standing there and waiting idly, no one can benefit from the cost of time, which is a greater waste. It means that the transaction cost of queuing is higher than that of using too much lubricant. That is why the gas station encourages customers to choose lubricant. However, different people have different cost of time, so there are still a considerable number of people who think queuing is a better choice. In short, under the constraints faced by the gas station, the transaction cost caused by price control for petroleum has been kept as low as possible, but it is still obviously higher than that in the house leasing market.

In fact, neither does the story in the house leasing have a happy ending. When government finds that contract fee is used to invalidate the rent control, it will come up with various supplementary policies to prohibit contract fee. Under this new constraint, the landlord has to think of new countermeasures which will definitely increase transaction cost, maybe to be even be much higher than that caused by tie sale of petroleum and lubricants in the price control for petroleum. Thus, although government regulation is always invalid, it does not mean that it has no adverse impact on society which refers to an increase in transaction cost.

There are two reasons for government to implement "evil policy": one is ignorance, the other is impudence. On the one hand, government may neither know the regulation will be only invalid, nor that if it tries to strengthen the regulation, it will only cause an increase in transaction cost. On the other hand, government is only a virtual concept, and it is government officials who are really alive, and they are also human and self-interested. They implement price control not because they are ignorant, nor because they really want help the poor (i.e., their claimed purpose is in fact just a pretense), but because regulation gives them power. When there is regulation, it is necessary to establish an organization (such as the Housing Department for rent control) to increase government spending which can feed or even fatten many government officials. In Lecture 2, it has been reminded that economics is a social science, which causes much trouble compared with a natural science: one is ignorance which leads to wrong application of economics, but the biggest trouble is that economists are also human and self-interested, and will deliberately distort economic analysis to deceive the public and benefit themselves. Those in favor of price control are mostly the laymen of economics who are ignorant, but some are economists. They have studied economics and still support price control, few of whom may be as ignorant as the laymen and should retake the course of economic, but I am afraid most of whom are imprudent.

Some may sympathize with the poor, and ask, "Without price control, what can the poor who cannot afford the price in the market do?" Firstly, from the above analysis, it is clear that price control cannot help the poor, but will only increase the transaction cost and make the whole society poorer.

Secondly, as long as the transaction cost is low enough, the market can satisfy any needs of all kinds of people (including the poor, of course). For example, many live under the same roof and share the rent, or a house is divided into several small single rooms which are rented separately, or even just a bed is rented out. The market will always supply goods of different quality according to different capacities of paying, and satisfy all kinds of demands. There is a saying that "big chickens do not eat small rice", which means that big businessmen disdain to earn small money and therefore disdain to serve small customers. However, in reality, big chickens eat big rice, small chickens eat small rice, and even broken rice is eaten by worms. As long as there is rice, no matter how small it is, there will always be someone to eat it. As long as there is money, no matter how small it is, there will always be someone to earn it. Therefore, as long as there is demand, there will be always supply and there will be always a market to take care of the demand. It is important for government not to interfere, misunderstanding that it can help the poor by preventing sellers from earning money which only results in strangling the supply for the demand.

Section 5: Minimum Wage Laws Cause Unemployment

As seen in the last section, the actual effect of price control that sets a price ceiling is so terrible, and what about price floor that sets a minimum price? It may seem contradictory for government to implement price floor to forbid low price while it implements price control that is claimed to help the poor who cannot afford the high price at the same time. Again it is the poor who are used as an excuse by government to implement price floor, because the price floor is mostly common implemented in the labor market and agriculture products market where the suppliers are labors and farmers who have always been regarded as the poor. Thus, minimum wage laws and agricultural price supports are the most typical price floors.

Let us look at the analysis of MSE first. Fig.9-4 shows the consequence of price floor. The equilibrium price in the market is P^* , but government imposes price floor to inflate it to P'' . At such a high price, buyers are reluctant to buy, so the quantity demanded is only Q_d'' . However, sellers are very willing to supply, so the quantity supplied is as high as Q_s'' . There is oversupply in the market, which is called "surplus". As a result, only the quantity of Q_d'' can be sold out. In other words, the consequence of price floor is fewer sellers' goods are sold out.

There is the same problem in this analysis of MSE as that for price control. The quantity demanded and supplied are both intentions, so the so-called "surplus" is not a fact which cannot be tested. Only based on specific constraints can be there meaningful economic analysis. Fortunately, price floor is generally only imposed in the labor market and agriculture products market, so let us analyze them one by one.

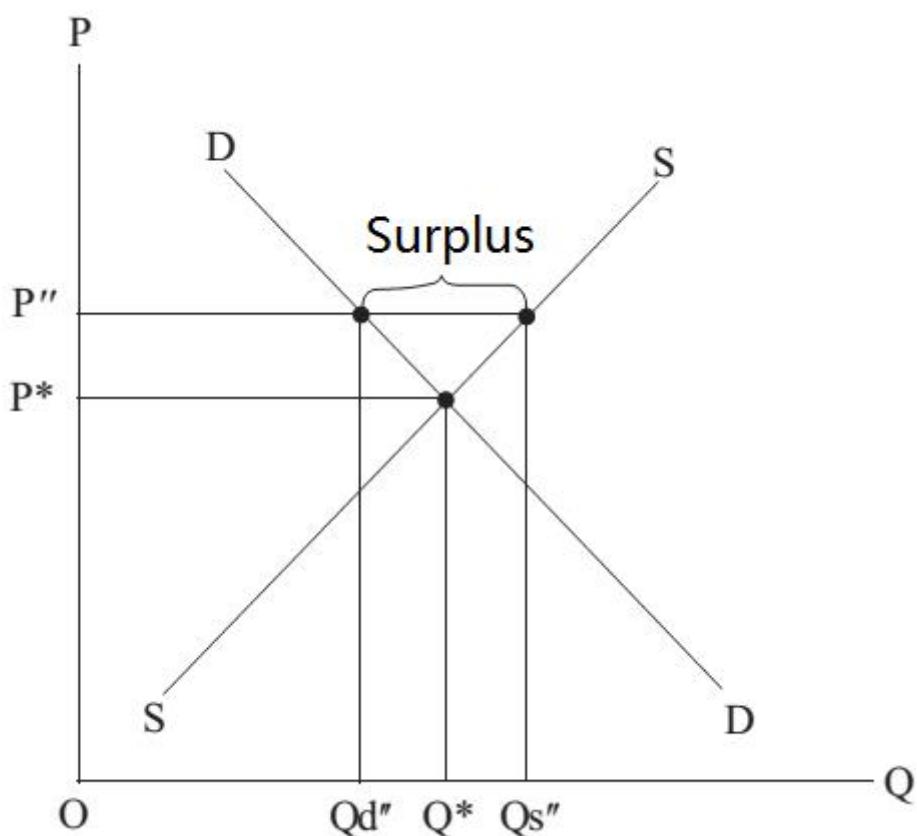


Figure 9-4

Minimum wage laws are generally analyzed in the part about factor markets in economic textbooks, but the relevant analysis actually applies the TSD here. If we specifically designate the variable of price as the wage rate, and the variable of quantity as the employment of labor force, we can easily change Fig.9-4 to Fig.9-5. It can be seen clearly in Fig.9-5 that when minimum wage laws set the minimum wage of W' above the equilibrium wage of W^* , there will be unemployment of $L_d L_s$. There are two sources for the unemployment of $L_d L_s$: one is that the rise in the wage rate causes the boss to dismiss the workers of $L_d L^*$, which means the quantity demanded for labor falls; the other is that the rise in the wage rate causes those who did not work before also come out to look for jobs, which means the quantity supplied of labor rises, so increase the unemployment of $L^* L_s$.

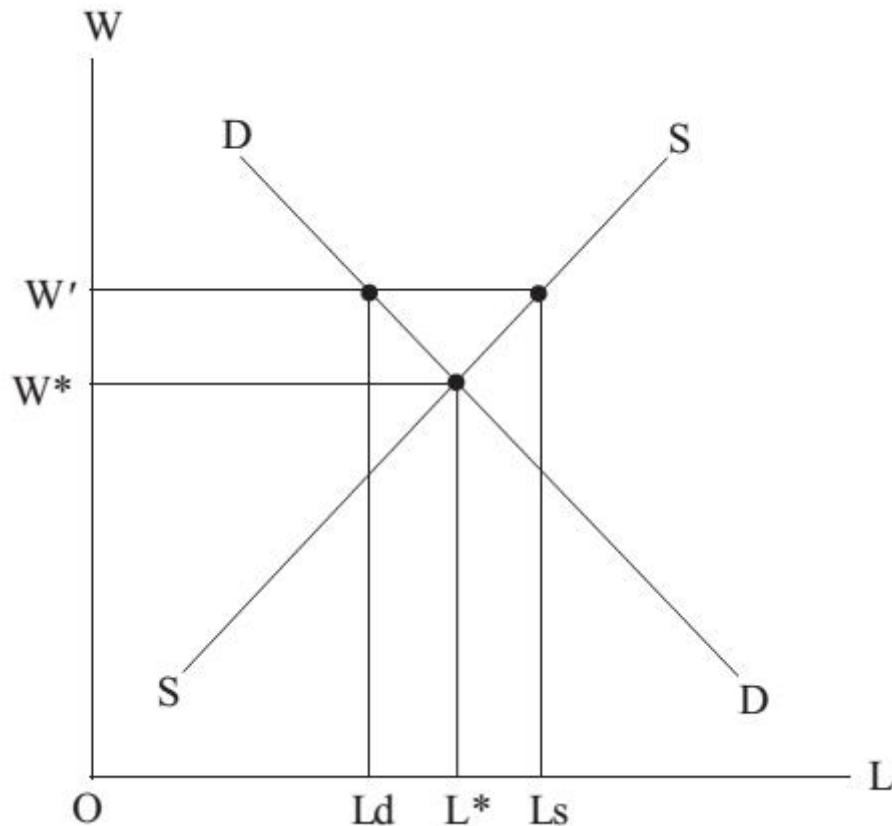


Figure 9-5

Yes, there is only one reason for unemployment: the wage rate is too high! It may sound hard to accept, but it is ruthless true. As long as the wage rate is low enough, a person with normal working ability can always have a job. He is unemployed just because he refused to accept the job that he thinks the wage is too low.

Unemployment is generally regarded as an issue of macroeconomics, but it is necessary to do some basic analysis here. Firstly, the definition of unemployment is problematic. What is unemployment? In macroeconomics, there is no clear and testable definition of unemployment. If those who do not have a job are unemployed, then a housewife who does not go out to work is unemployed, which is ridiculous. A rich kid needs not work and can still live well, so he idles around all day. Is he unemployed? Obviously, this kind of unemployment include housewives and idle rich kids is irrelevant with the macroeconomic problem government should care about.

Therefore, macroeconomists divide unemployment into voluntary one and involuntary one. Voluntary unemployment refers to the cases such as housewives and rich kids mentioned above, which is not a problem to solve, while involuntary unemployment is. However, again what is involuntary unemployment? Here is a job of working 24 hours a day with the wage of only RMB1. It is hard to imagine anyone would accept this job, but those who refuse it are "voluntarily" unemployed literally.

Again, macroeconomists are busy in revising the definition of involuntary unemployment as that the labor force cannot find a job at the wage level he is willing to accept. However, this definition is still problematic. Some say he is willing to accept a job with the wage of RMB100 trillion, and there cannot be such a high-paid job in this world. Is he involuntary unemployment? To define unemployment as voluntary or involuntary means to use psychological factors to explain the unemployment phenomenon, which is non-scientific as discussed in Lecture 8 of "the Law of Demand" about the prohibition of using psychological factors for explanation in economics. In other words, in macroeconomics there has been no scientific definition of unemployment.

However, if seen from the perspective of ordinary goods, it is not such a difficult problem. That a labor is unemployed is equivalent to that a good cannot be sold out. As long as it is somewhat useful, it is impossible to be unable to be sold out by cutting the price. Even scrap metal, old books and newspapers can be sold out at a price low enough. Thus, the price is too high relative to its use, which is the only reason why it cannot be sold out.

It is true for labor force. As can be seen from fig.9-5, the minimum wage of W' is too high compared with the equilibrium wage of W , *which is the only reason for the unemployment of L_dL_s . If wage is allowed to drop to W , on the one hand, the number of L^*L_s will choose to withdraw from the labor market, which means they will become housewives or rich kids voluntarily unemployed. They are those who ask for relatively higher wage, so if the wage offered by the market cannot meet their requirements, they will rather choose not to work. Take a housewife as an example, she will choose to work as long as the wage she gets from the market is so high that she can afford to hire a nanny to take care of the children and do housework for her. It is logically similar to a rich kid. If the market pays him tens or even hundreds of millions of dollars, why does he not come out to work? Just because generally the market will not offer him such a high wage, he chooses to be unemployed voluntarily.*

On the other hand, employers will increase their quantity demanded of L_dL^* for labor force. When the minimum wage is too high, employers will not employ these people because they are not worth the price. As pointed out earlier, the essence of price floor is trying to infringe buyer's right to income right, which is foreseen by buyers before the transaction, so buyers simply choose not to buy to avoid the infringement. So what is actually infringed is the right to transfer of sellers. Yes, government can impose a higher price, but it cannot force sellers to buy something that they think not worth that high price. As a result, the good with price floor cannot be sold out, which is precisely unemployment in the labor market. In other words, the essence of minimum wage law is an infringement of the right to transfer of the labor force, not allowing workers to rent out their labors at a price lower than the regulated minimum wage, and it actually deprives workers of their right to work!

The objective effect of government's policy is once again contrary to its stated good purpose: the minimum wage laws claim to help the poor obtain higher incomes, but in fact make them unemployed and lose the meager incomes they could have obtained. The minimum wage laws claim to protect the interests of workers, but in fact infringe the PPRs of the labor ability which is the only asset a worker may have. Moreover, it is low-wage people who suffer most, because the minimum wage laws only have effective impact on those whose equilibrium wage rates are lower than the regulated rate. The following reading material "Do Minimum-Wage Laws Reduce Employment?" has listed the results of the relevant studies: from rich and large countries like the US, to poor countries like Kenya, to small countries like Singapore, all experiences show that minimum wage laws have reduced employment, and it is low-wage people who suffer most from unemployment caused by minimum wage laws. Faced with such eloquent facts, how can anyone doubt about the conclusion?

READING MATERIAL: DO MINIMUM-WAGE LAWS REDUCE EMPLOYMENT? 1

Economists, at least up to recently, have believed that in most labor markets the evidence supports the competitive model. A minimum wage, set higher than the previous equilibrium level, would be expected to reduce employment among the affected low-wage groups. The table here indicates some of the results of a study by Donald Deere, Kevin M. Murphy, and Finis Welch dealing with minimum wage increases legislated in 1990 and 1991.

Percent employment reduction after 1990-1 minimum wage Increases

Age group	% low wage	% employment change
Men		
15-19	44.5	-15.6
20-24	14.2	-5.7
25-64	3.3	-2.4
65-69	14.0	-4.2
Women		
15-19	51.8	-13.0
20-24	19.0	-4.2
25-64	8.8	-0.3
65-69	21.0	+3.1

Source: Extracted from Deere, Murphy, and Welch, p. 48.

The table shows that age-sex groups with the highest proportions of low-wage workers – most notably the 15–19 age groups – generally suffered the largest proportionate reductions in employment. This is exactly what the competitive model would predict. (A possible exception was women 65–69, but only a small percentage of this group was employed to begin with.) The investigators also found large employment drops across other demographic groups with high proportions of low-wage workers, among them African-Americans and persons of Mexican descent.

Furthermore, the U.S. results confirming the competitive model have generally been supported by experiences with minimum wages in countries as different as Great Britain, Singapore, and Kenya.²

As an interesting exception, however, a study by David Card and Alan Krueger obtained results in conflict with the competitive model.³ The competitive theory predicts that an increase in the New Jersey minimum wage would reduce employment of low-wage workers in New Jersey relative to the neighboring state of Pennsylvania, but if anything the opposite was found.⁴

From this analysis we can again learn another philosophy: do not expect government to help you, it cannot not help you, but even harm you! Whether government is ignorant or impudent, the objective effect is that it cannot help you.

In fact, all macroeconomists (including Keynes) admit that the direct cause of unemployment is the too high wage. Various theories about unemployment theory in macroeconomics are just trying to explain why the wage is too high and why the market cannot reduce the wage. However, these macroeconomists are simply unwilling to face the reality that there are minimum wage laws and strong trade unions in western countries that block the wage to be adjusted downwards, thus directly nullifying the market's function of equalizing the supply and demand with price change. There are other complicated aspects of unemployment which will be explained in the later lecture.

There are still some key points that need to be discussed about minimum wage laws. Firstly, some argue that minimum wage laws will not cause unemployment. Suppose the enterprise need 100 persons for its operation without minimum wage laws. When a minimum wage law is imposed, according to which no one's wage can be less than what the minimum wage law stipulates, and the enterprise still need 100 persons to maintain its operation, so it has to employ all the labors as before, and no one will be fired.

It is so foolish to think so! There are many ways for enterprise to reduce the demand for labor, such as increasing labor intensity. During the financial tsunami of 2008, a reporter interviewed the security guard of a factory. He said the boss fired the cleaner and required him to undertake the cleaning job along with his original responsibility. Although he was not unemployed and got the same wage as before, he were responsible for two jobs, which was actually a disguised wage cut. If there were no minimum wage laws, the cleaner would keep his job with a wage cut, and the security guard would also keep his workload unchanged with a wage cut. However, with minimum wage laws hindering the wage cut, the cleaner who had the lowest wage loses his job and the security guard increases his workload as a disguised wage cut. What's more, why must the enterprise maintain its original scale of operation? It can downsize the operation, and so cut some jobs to avoid the increase of labor costs caused by minimum wage laws. In extreme cases, it can even go bankrupt and reduce the number of employment to zero!

Secondly, in reality, the immediate response of the enterprise affected by minimum wage laws is generally not to dismiss employees, but to reduce or even stop recruiting new employees, because it is more troublesome (i.e. higher transaction cost) for the former than the latter. In other words, it is not those who have jobs but those who are looking for jobs that bear the brunt of the impact of minimum wage laws, which means that graduates just leaving schools suffer most. Indeed, the unemployment rate of young people is relatively high. However, if the minimum wage continues to rise or the performance of economy becomes worse, the enterprise will increase labor intensity as a disguised wage cut as mentioned just now. Finally, some employees will be fired, and those who are least worth or whose equilibrium wages differ most from what minimum wage laws stipulate will be the first to be fired.

Thirdly, some will doubt if the market is free to determine the wage, those with the worst working ability may get so little income that they cannot make ends meet. What should we help them? It is similar to the worry that the poor cannot afford to rent a house without rent control, so the answer is also similar. According to the analysis above, minimum wage laws cannot help those with the worst working ability to get higher wages, but only make them unemployed and have not any income at all. In contrast, as long as it is not interfered by government's evil policies, with the minimum wage determined by the market, the poor is sure to be able to make ends meet, because the prices of a worker's products and his wage are related to each other. On the one hand, the later lecture on the income distribution will point out that a worker's wage is determined by the prices of his products instead of the boss who employs him. On the other hand, the wages of workers will in turn affect the prices of products. If the wages of workers are low, the prices of products will be low too, so the expenses needed by workers to make ends meet will fall. If the wages of workers are too low to make ends meet, there will not be enough purchasing power to support the prices of products. The rich will not consume a large amount of

products used for living, and will not compete with the poor for these products, so it is the poor's consumption that determined the prices of them. In other words, the prices will fall until the poor can afford to make ends meet. That is why the minimum wage determined by the market must be enough to make the poor survive.

1. This reading material is an excerpt from "Price Theory and Its Application", P369. [e](#)
2. Deepak Lal, "The Minimum Wage: No Way to Help the Poor," The Institute of Economic Affairs, Occasional Paper 95 (1995). [e](#)
3. David Card and Alan Krueger, "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," American Economic Review, v. 84 (September 1994). [e](#)
4. But see David Neumark and William Wascher, "The Effects of New Jersey's Minimum-Wage Increase on Fast-Food Employment: A Re-evaluation Using Payroll Records," National Bureau of Economic Research Working Paper 5224 (August 1995). [e](#)

Section 6: The Consequence of Agricultural Price Supports

Although agricultural price supports and minimum wage laws both belong to price floors, their constraints are different, so their phenomena are also different. Government can force enterprises to pay high wages through minimum wage laws, but cannot force consumers to buy agricultural products at high prices. In order to implement agricultural price supports, government has to purchase all the agricultural products that are so expensive that no one will buy in the market. Imagine what would happen if this kind of government purchase were used in the labor market. Government does not enact minimum wage laws, but establishes state-owned enterprises to recruit workers with higher wages than those offered by private enterprises. Obviously, people will flock to apply for state-owned enterprises, which forces private enterprises to follow and raise the wages to the same level, or else they cannot recruit any workers.

There are two direct effects of agricultural price supports: firstly, government must pay for a large number of so-called "surplus" agricultural products; secondly, agricultural products in the market are extremely "scarce", so the prices are extremely high. Government can solve the latter problem by supplying the agricultural products it buys at a high price to the market at a low price or for free. As a result, there are phenomena such as free lunch for students and food aid used as a diplomatic tool in the US.

By the way, is free lunch really free? Do not forget that government is not an enterprise, so it does not produce to earn income, and its income comes from taxes which are collected from taxpayers. The so-called free lunch provided by government is in fact indirectly paid by taxpayers. Taxpayers are forced to pay no matter whether they have children in schools or not, no matter whether they are willing to pay for the lunch or not. If one pays directly for the lunch, he will assure the quality is good enough for the price they pay. He may be willing to choose better or worse (with lower price) quality than that of the free lunch. Now government has made a choice on his behalf. Is it possible that government cares more about his children and makes a better choice than he does? For the suppliers of the free lunch, if customers pay directly, the saying that "customer is god" is not an empty moral sermon, but the truth supported by money. However, if customers pay indirectly through government, it is government that is god. The suppliers only care about how to please government (strictly speaking, government officials), rather than how to provide the free lunch good and cheap to please customers.

READING MATERIAL: AGRICULTURAL PRICE SUPPORTS 1

Since the 1930s the U.S. government has attempted to maintain "parity" prices for agricultural products. Parity means the relationship between agricultural and nonagricultural prices that obtained during the years 1910 and 1914, a period of farm prosperity. Throughout the 1950s and 1960s, high parity price floors were maintained by government purchases. A federal agency, the Commodity Credit Corporation (CCC), would buy any otherwise unsold quantities of supported crops at the support level usually at 90% of the parity price.

The "surpluses" purchased by the CCC were mostly stored; the federal government hoped to sell them in years of small harvests. But the amount produced almost every year was greater than the amount consumers would purchase at the artificially high supported prices. By 1960 the CCC held in storage as much wheat as that year's entire crop.

To reduce the cost of holding such huge stores, Congress passed food stamp and school lunch programs subsidizing food prices for some consumers. A variety of "supply management" programs were introduced to reduce quantities produced. In addition, the price support levels were reduced. The table shows that, accordingly, in the 1960s government purchases were much smaller than in the previous decade.

Yearly acquisitions of three supported crops by the commodity credit corporation, selected years

Grain	Sorghum	Corn	Wheat
1953	40.9	422.3	486.1
54	110.1	250.6	391.6
55	92.6	408.9	276.7
56	32.5	477.4	148.4
57	279.6	268.1	193.5
58	258	266.6	511
1963	125.1	17.9	85.1
64	66.8	29.1	86.9
65	85	11.2	17.2
66	0.3	12.4	12.4
67	9.1	191	90
68	13.7	34.4	182.9

Source: Commodity Credit Corporation charts, November 1972, pp. 49, 75, 115.

Over the years, political support for the policy of buying surpluses has declined. Instead, a variety of other techniques for maintaining farm prices have been increasingly used, the details varying from commodity to commodity. Generally speaking, these restrict either the number of acres that can be planted or the amount of product that can be sold. Such methods of "supply management" have been supplemented by other programs aimed at maintaining farm incomes, among them direct payments to farmers for acres not planted or for crops not sold.

Government forces enterprises to bear the economic burden of minimum wage laws, while it has to bear the economic burden of agricultural price supports by itself. Therefore, government has been trying to get out of this trap it accidentally fell into, but hardly succeeds. The following reading material of "Agricultural Price Supports" shows how the US government has struggled for decades: firstly to reduce the level of price supports, then to restrict farmers' planting areas in an attempt to reduce agricultural outputs. However, under the stimulus that the supported prices are still higher than the equilibrium prices, farmers have also played the trick of going around the restrictions: when government forbids them to increase the planting areas, they will use all kinds of means to increase the outputs per unit area. Finally, the U.S. government has to directly limit the amount of agricultural products that farmers can sell at the high prices set by price supports.

In fact, for European and US governments, there are mainly two mountains that have overwhelmed their finances and plunged them into sovereign debt crisis, one of which is precisely the fiscal expenditures on agricultural price supports. The other one is social security on retirement pension, which will be analyzed later in this book.

1. This reading material is an excerpt from "Price Theory and Its Application", P48-49. [e](#)

Lecture 10: Consumer Theory

In the textbook of MSE, it is usually consumer theory after the TSD.

However, with the law of demand as an axiom, consumer theory is actually redundant, because it tries to derive the law of demand as a theory. In consumer theory, there are two axioms (comparison axiom and transitivity axiom), a concept of "utility" and an analysis by using indifferent curves, with which a price expansion path due to price changes are drawn to deduce a demand curve. Nevertheless, with the disturbance of the Giffen good, this demand curve cannot be sure to slope downwards, so it is necessary to assume that there is no Giffen good. In short, if the explanatory power of consumer theory is as strong as the law of demand, there must be much more contents, which make the theory extremely complicated.

According to "Occam's Razor" in the methodology of science introduced in Lecture 8, the law of demand is obviously better than consumer theory. What's more, the concept of utility and indifferent curve, on which consumer theory depends deeply, are both intentions. When applying them to explain phenomena, one must first try to find observable facts to replace them, which also make the applications much more complicated. There are traps of tautology easily to fall into.

Furthermore, in practical application, the explanatory power of the law of demand law is much stronger than consumer theory. As pointed out in the last section of Lecture 8, if the variable of "price" is interpreted broadly as "cost", not only the behaviors of consumers but also of producers can be explained. If it is further interpreted as "constraint", the law of demand can actually explain all human's behaviors. That is why the law of demand is not just an ordinary theory but an axiom in economics.

In fact, there is no need for practitioners to learn consumer theory. Then why do I still write this lecture? Because this theory is strong in logic, and it is good for training the logical thinking. However, if you want to explain phenomena, do not expect consumer theory to be helpful, and the law of demand is much simpler and more useful.

Section 1: Two Axioms of Consumer Theory

There are two axioms in consumer theory, both of which are postulates about human's preference (the same as "taste"). One is the axiom of comparison: a person can compare any two baskets A and B of commodities. Such a comparison must lead to one of the three following results: he or she (i) prefers basket A over B, or (ii) prefers basket B over A, or (iii) is indifferent between A and B.

The other is the axiom of transitivity: consider any three baskets A, B, and C. If a consumer prefers A to B, and also prefers B to C, he or she must prefer A to C. Similarly, a person who is indifferent between A and B, and is also indifferent between B and C, must be indifferent between A and C.

These two axioms taken together lead to the proposition of rank ordering of preference: a consumer can consistently rank all baskets of commodities in order of preference. This ranking is called "the preference function."

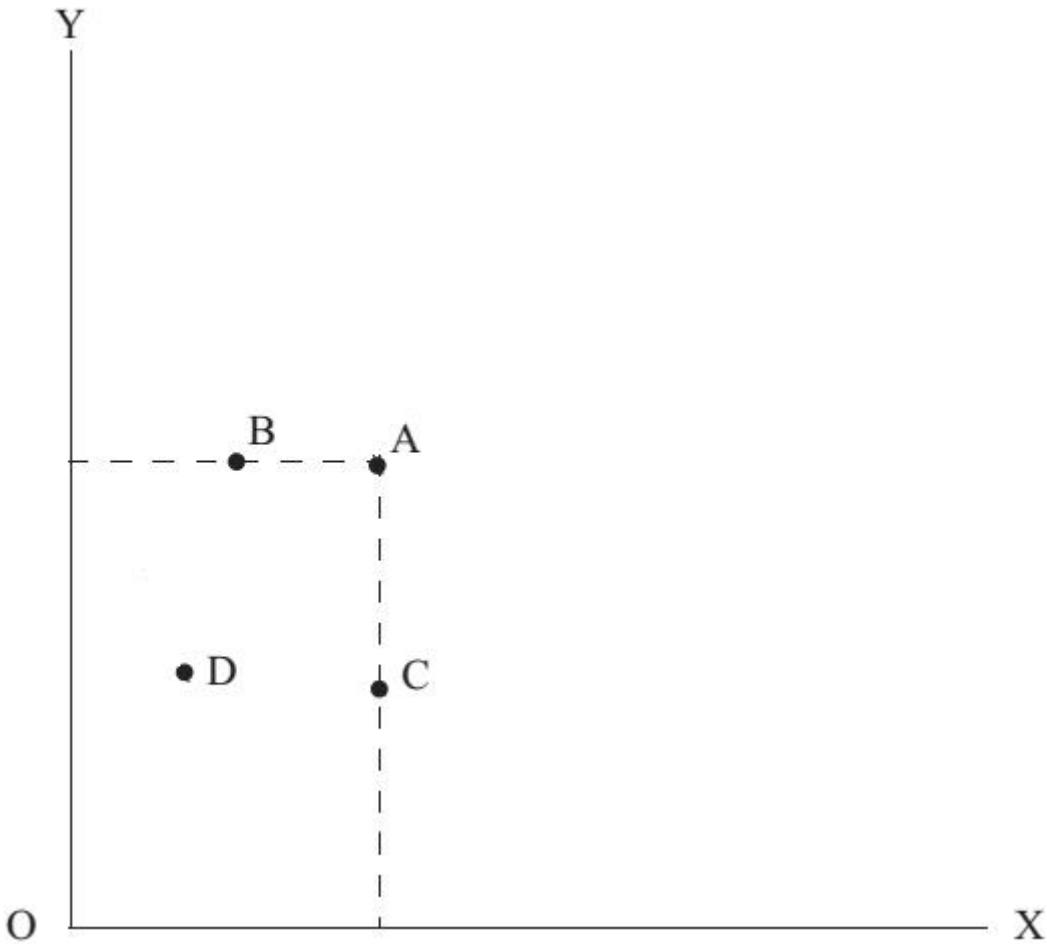


Figure 10-1

Based on the above axioms, proposition and the definition of "goods" introduced in Lecture 4, we can draw the Fig.10-1. The x-axis and y-axis respectively represent the quantity of good X and good Y, and any point in the rectangular coordinate is a basket of different quantities of these two goods. For example, the x-coordinate of point A represents the quantity of good X in the basket A, while the y-coordinate is the quantity of good Y. According to the above axioms, laws and the definition of "goods", we can judge that point A is preferred to D, because it contains more of X

and Y than D does. It is the preference ranking for the two goods according to the proposition of rank ordering of preference.

However, here is the question: which is better, point B or point C? How to rank their orders? The quantity of Y at point B is more than that at point C, while the quantity of X at point B is less than that at point C. The former means B is preferred to C, but the latter means C is preferred to B, which implies that only with the above axioms, proposition and the definitions of "goods", the constraints are not enough to infer human's behavior of choosing which basket of goods between B and C.

The answer to this question is the analysis tool of "indifference curve" which will be derived in the later section.

Section 2: Use Value VS Utility — Methodology of Science (3): What is measure in science?

Before deriving indifference curve, a concept of "utility" should be discussed first. What is utility? Chinese students all have learned Marxist economics in middle school. There is a concept of "use value" in Marxist economics, which is not invented by Marx, but by Adam Smith and is widely used by classical economists. "Use value" is actually a very good concept, but it was invented along with the "theory of labor value" which was refuted by facts and abandoned now, so it is also abandoned and replaced by the concept of "utility" which is worse than it.

The concept of "utility" was invented by Jeremy Bentham (1748-1832), a British philosopher, who was famous for his philosophy of "utilitarianism". This concept was introduced into economics by the economists who initiated the marginal revolution, and replaced the concept of "use value". However, economists have completely changed the original meaning of utility which meant the pursuit of happiness and avoidance of pain in utilitarianism advocated by Bentham.

Now in economics, utility is a measurement of human's preference for goods, which means it is used to measure the psychological factor of preference, so it is not a fact, and if directly applied to explain phenomena there will be the mistake of tautology. In order to avoid the trap of tautology, we have to look for the objective constraints that cause the psychological factor of preference, which means we must ask: why does a person prefer a specific good? The answer is because the good can meet some of his needs, which simply means that the good has "use value" for him! In other words, "use value" is directly related to objective constraints and is a fact, while "utility" is directly related to psychological factor of preference and is not a fact. Compared with "use value", when "utility" is applied in explanation, it is either easy to fall into the trap of tautology, or we need to look for objective constraints to replace the psychological factors, which is too long-winded and against "Occam's Razor" in methodology of science, so it is not a good concept. However, utility is easy to use mathematics to represent, so it is widely used in academic papers of economics. How much there is empty economic paper just filled with utility analysis in the form of mathematical equations nowadays!

Therefore, we advocate abandoning the concept of "utility" which is good-looking but empty, and returning to the concept of "use value" which is simple but objective.

However, in the concept of "use value" advocated by Adam Smith (and Marx), there is a major mistake which must be corrected. In Smith's view, the same good has the same use value for different people. In fact, the use value of the same good can be completely different for different people, times and places. For example, when I am a teacher, and I keep talking all the class, so I need to take water which is of high use value to me. But for the students who just sit and listen most of the time, water is not absolutely necessary and of low use value. Therefore, the use value of water is different for different people, which does not mean it is psychological.

Furthermore, the use value of water can be different at different times and in different places. Suppose one is in the desert without water, and another one appears in front of him with a bottle of water. How much would he like to spend to buy the water? The answer is: he would be willing to give all his money in exchange. The use value of the water now is so high for him under such circumstances that it is equivalent to the value of life!

By the way, there is also a concept of "exchange value" in Marxist economics (actually in classical economics) which is distinguished from price. However this distinguish is meaningless, and exchange value is the same as price. Of course, there is relationship between price (exchange value) and use value (utility), but not the kind of relationship advocated by Marxist economics, which will be discussed in the later lecture.

Early economists thought utility could be measured quantitatively, which means it was cardinal magnitude. However, later they found that utility could only be ordinal magnitude. Cardinal magnitude is one, two, three ..., which can be used for four fundamental operations of addition, subtraction, multiplication and division. For example, the length is cardinal magnitude, so it can be calculated that a rope of 3 meters is 2 meters longer than a rope of 1 meter, and the two ropes are 4 meters long in total when connected. And ordinal magnitude is the first, second, third ..., which can only be sorted, but cannot be used for four fundamental operations. What is the first plus the third? It cannot be the fourth.

Since ordinal magnitudes can only be sorted but not be used for four fundamental operations, cardinal magnitudes are often used to replace them. For example, in the beauty contests mentioned above, the judges will give scores to each contestant. The scores are cardinal magnitudes which can not only be sorted, but also be used for four fundamental operations. Suppose the champion's score is 90 points, the second runner-up's is 85 points and the third runner-up's is 70 points, the gap between the champion and the second runner-up is smaller than that between the second runner-up and the third runner-up. However, these scores are only for the convenience of summing up the contestants' performances in different contest sections to get a final result. So they do not really reflect the gap between the different runners-up. When an ordinal magnitude is replaced with a cardinal magnitude for convenience of operation, we must be very careful not to mistakenly include the properties that the ordinal magnitude does not have but are included in the cardinal magnitude. That is why the utility theory is full of traps. In addition to the aforementioned problem that utility is intention and cannot be directly used to explain phenomena, there is also risk of making mistakes when economists use mathematics to express utility. Do not forget that utility is ordinal magnitude rather than cardinal one, so the four fundamental operations may not be appropriate in analysis.

Here is the third part of the methodology of science in this book: what is measure in science? As for "measurement", those who do not understand MS generally have two interrelated misunderstandings: firstly, they think that only what is measurable is scientific. Secondly, they think there is only cardinal measurement. The combination of these two misunderstandings leads to the wrong conclusion that only "quantitative analysis" (as opposed to "qualitative analysis") is accurate or "measurable", and is scientific. Marx made such a mistake, saying that "only by making full use of mathematics can any science reach a perfect state", which is totally wrong from the perspective of MS. As has been explained in Lecture 2, it is "refutability" that is the essential characteristic of science, while mathematics is not science, which is only a tool that can but not must be used to assist in the logical reasoning in the second stage of science. Further, mathematics is not equivalent to measurement which is statistics! What's more, there is not only quantitative analysis or cardinal measurement. There have been three mistakes of disguised replacement of concept here: to replace the concept of science with mathematics, to replace the concept of mathematics with statistics, and to replace the concept of statistics with quantitative analysis, which results in the replacement of truth with accuracy appeared by quantitative analysis. It is precisely the so-called "accurately false".

In recent years, there have been such accurately false developments in many social sciences, which is exactly economics that has firstly abused mathematics. Economics is closest to natural science in social science because Adam Smith, the founder of economics, found the postulate of "self-interest" as the first stage of science, and Marshall introduced mathematics into economics as a tool to assist logical reasoning in the second stage of science. It is clear that all Marshall did

with mathematics were qualitative analysis, so he knew very well that mathematics is not equal to cardinal measurement. It is on the third stage of fact tests that economics unknowingly went astray. Economists who do not understand MS see that in the second stage mathematics is widely used, just like natural science, they mistakenly think that in the third stage economics also needs quantitative analysis, just like natural science. Science requires fact tests, but it has never required that only data are facts. Natural science can do experiments in laboratories, so data are generally facts, and most of them are cardinal magnitudes. In contrast, social science is the study on human beings and cannot do experiments in laboratories, which will be explained in detail in the later lecture of "concept of cost", so data are not necessarily equal to facts. What's more, the data suitable for describing human behaviors are generally ordinal magnitudes instead of cardinal ones. Take the beauty contests mentioned above as an example. The champion got the highest score of 90 points, the runner-up 85 points, and the third runner-up 70 points. These three cardinal magnitudes are only meaningful in ranking orders.

Nowadays, in many social sciences, econometric (statistical) analysis is widely used as fact tests to show that they are scientific. However, social science cannot do experiments, and there are too few social statistical data, so many studiers use "questionnaire survey" to obtain enough data. In questionnaires, there are full of such question: "do you think something is (a) very important; (b) important; (c) less important; (d) completely unimportant," and then arbitrarily assign each option such 10 points for (a) and 7 points for item (b) ... and then sum up these cardinal magnitudes which are only meaning full in sorting. If only the summation results are sorted to show whether most think that something is the first or the second important just like the winners order in beauty contests, it makes some sense. After all, it is still back to the nature of ordinal magnitude, although the logical reliability of the summation in the intermediate stage is very doubtful. It is worse that some studiers directly put the data obtained from the summation of the ordinal magnitudes so suspiciously into the regression equation, and make a "student's t test" or something like that, then claim that the theory has been tested ... They ignore the ordinal nature of the data which are used for four fundamental operations. No matter scientists respecting facts or mathematicians respecting strict logics will never admit what they do is scientific or logic. However, there are full of such so-called test which is neither scientific nor mathematical in all kinds of academic publications in social sciences. What is the difference between them and pseudo-science by cheaters?

In the third stage of science, to be "roughly true" is much preferred to "accurately false". Precision is not equal to truth, and accurate mistakes are pseudo-science in essence, which is worse than non-science! Non-science, such as religious belief, ethics, mathematical logics, also has its significance in some aspects, while pseudo-science is pure fraud. The "roughly true" conclusions come from ordinal measurement which seems not very accurate from the perspective of cardinality, but is logically reliable. As for how to use ordinal measurement, the "marginal analysis" introduced in lecture 7 is a powerful tool.

Section 3: Indifference Curve

With the concept of use value (utility), we can go back to solve the problem of ranking the different baskets of goods. Looking back at the fig.10-1 in section 1, each point in the rectangular coordinates is a unique basket of good X and Y, and one can rank these points (baskets of goods) by their use values to him. If he thinks that the use values of the baskets represented by points of B and C are the same, a curve called "indifference curve" can be drawn to through them. Similarly, two other indifference curves can be drawn through point A and D respectively. So the fig.10-1 in section 1 is turned to fig.10-2 here.

Indifference curve has four crucial properties.

(1) Difference curve has negative slope, which means it is a curve sloping downwards. As shown in Fig.10-2, since point A is preferred to D, a curve passing through these two points and sloping upwards cannot be an indifference curve. Because both X and Y are goods of which more is preferred to less, it is necessary to decrease the quantity of the other good, when the quantity of one good increases and the use value of the basket must be kept unchanged. In other words, the points with the same use value must move along a curve where the quantity of one good increases and that of the other good decreases. It means the slope of this curve must be negative.

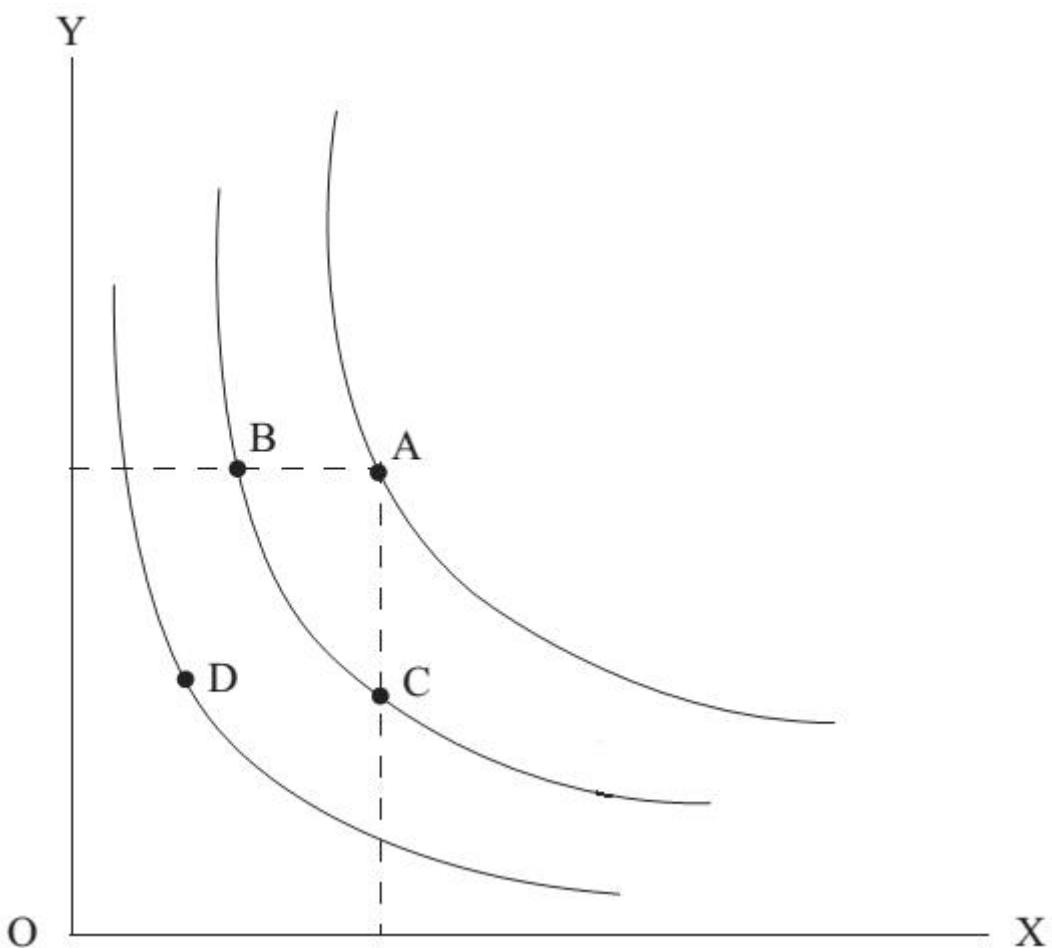


Figure 10-2

(2) Indifference curves never intersect, which can be proved by reduction in mathematics. Assume tentatively that two indifference curves such as U₁ and U₂ do actually intersect at point A in Fig.10-3. Points A and C are both on U₁, which means the consumer is indifferent between them. Points A and B are both on U₂, which means the consumer is indifferent between them. By the axiom of transitivity, the consumer must be indifferent between B and C, which means they

should be on the same indifference curve, but in fact they are not. So the initial assumption is invalid, and indifference curves cannot intersect.

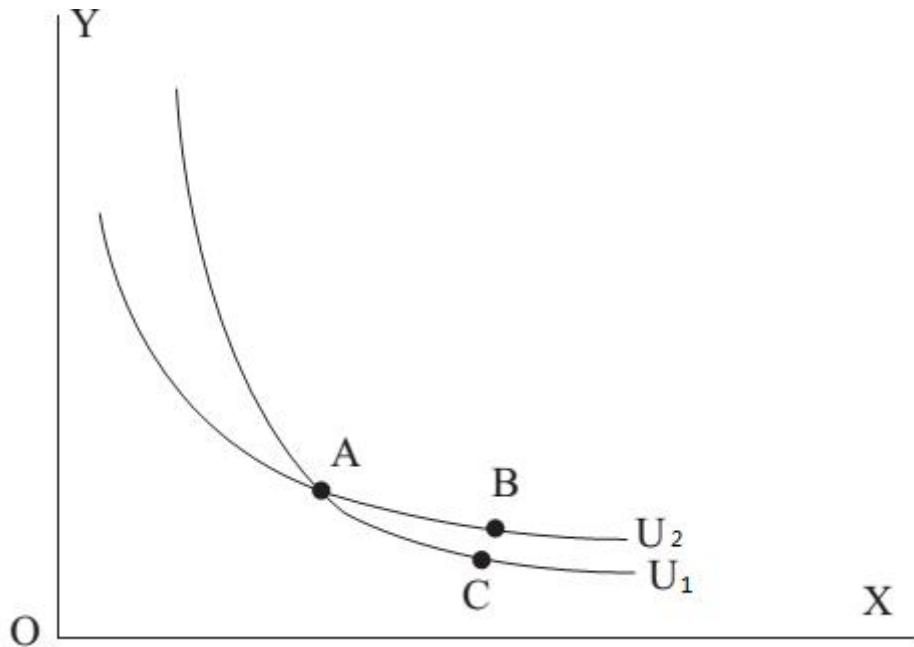


Figure 10-3

(3) The use values of points on the same indifference curve are the same, but the use values of points on different indifference curves is certainly different. How to determine which one has higher use value? The answer is: the indifference curve farther away from the origin has higher use value. The proof is as followed: As shown in fig.10-2, it is obvious that the indifference curve through point A is farther away from the origin than the curve through points B and C, while the indifference curve through point D is closer to the origin than the curve through points B and C. Because point A is preferred to points B and C, the indifference curve through A, which is farther away from the origin, has a higher use value. In contrast, both points B and C are preferred to point D, so the indifference curve through B and C, which is closer to the origin, has lower use value.

The use values of points on the same indifference curve are the same, but the use values of points on different indifference curves is certainly different. How to determine which one has higher use value? The answer is: the indifference curve farther away from the origin has higher use value. The proof is as followed: As shown in fig.10-2, it is obvious that the indifference curve through point A is farther away from the origin than the curve through points B and C, while the indifference curve through point D is closer to the origin than the curve through points B and C. Because point A is preferred to points B and C, the indifference curve through A, which is farther away from the origin, has a higher use value. In contrast, both points B and C are preferred to point D, so the indifference curve through B and C, which is closer to the origin, has lower use value.

(4) Indifference curve is convex to the origin. According to (1), indifference curve must be sloping downwards. The indifference curves in left and right panels of Fig.10-4 are both sloping downwards, but the shapes are different. The one in the left panel is convex to the origin, while the one in the right panel is concave to the origin. The one in the left panel is based on the well-established empirical principle of "diversity in consumption", while the one in the right panel comes from "unity in consumption".

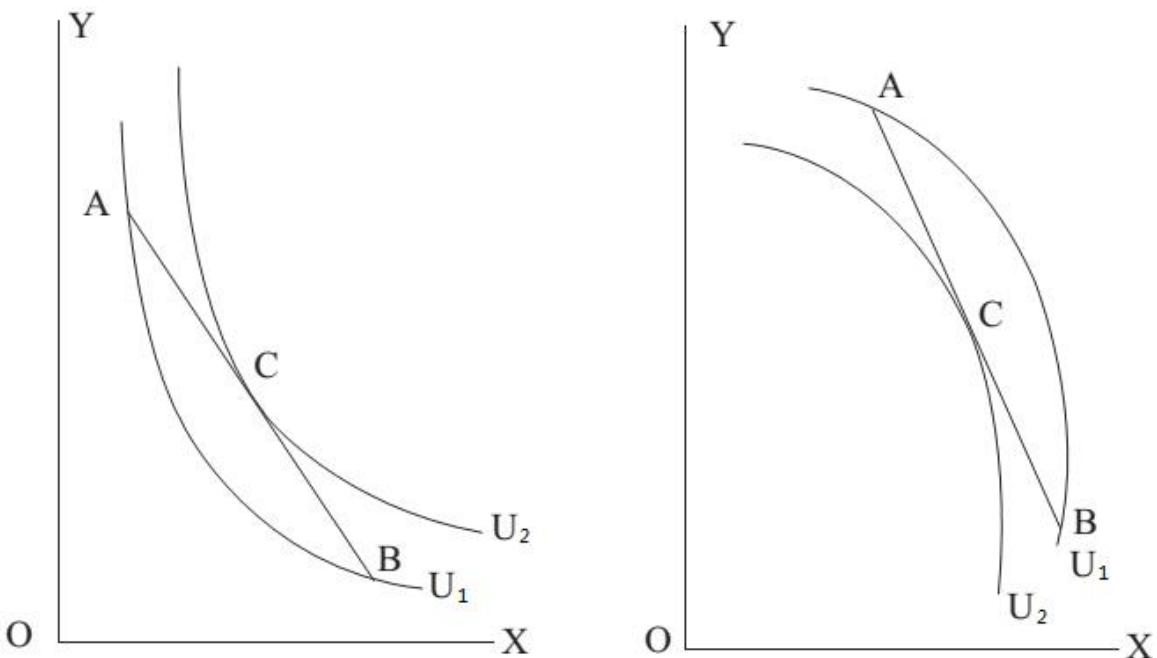


Figure 10-4

We can choose two points A and B on the indifference curve of U_1 in both panels of Fig.10-4, and connect these two points with a straight line, then take the middle point C through which draw another indifference curve of U_2 tangent to the line AB. It is clear that use value represented by U_2 is higher than U_1 in the left panel, while lower in the right panel. Obviously, the quantity of two goods in point C is more balanced than that in points A and B. The straight line connecting points A and B is (relative) price line for the two goods, which will be explained in the later section. The left panel means that at a certain price, people think it is better to consume the two goods in a balanced way (represented by point C) than in an extreme way (represented by point A with too much Y and point B with too much X). In contrast, the right panel means that at a certain price, people think it is better to consume the two goods in an extreme way than in a balanced way. The left panel is supported by people's consumption behaviors in reality, so indifference curve should be convex to the origin.

In Lecture 8, complements and substitutes have been introduced, and their indifference curves are drawn in the left and right panels of Fig.10-5 respectively. The indifference curve of perfect substitutes is a straight line, while that of perfect complements is a right angle. Most of goods in reality are between these two extremes, so the indifference curve of two goods is generally a curve convex to the origin.

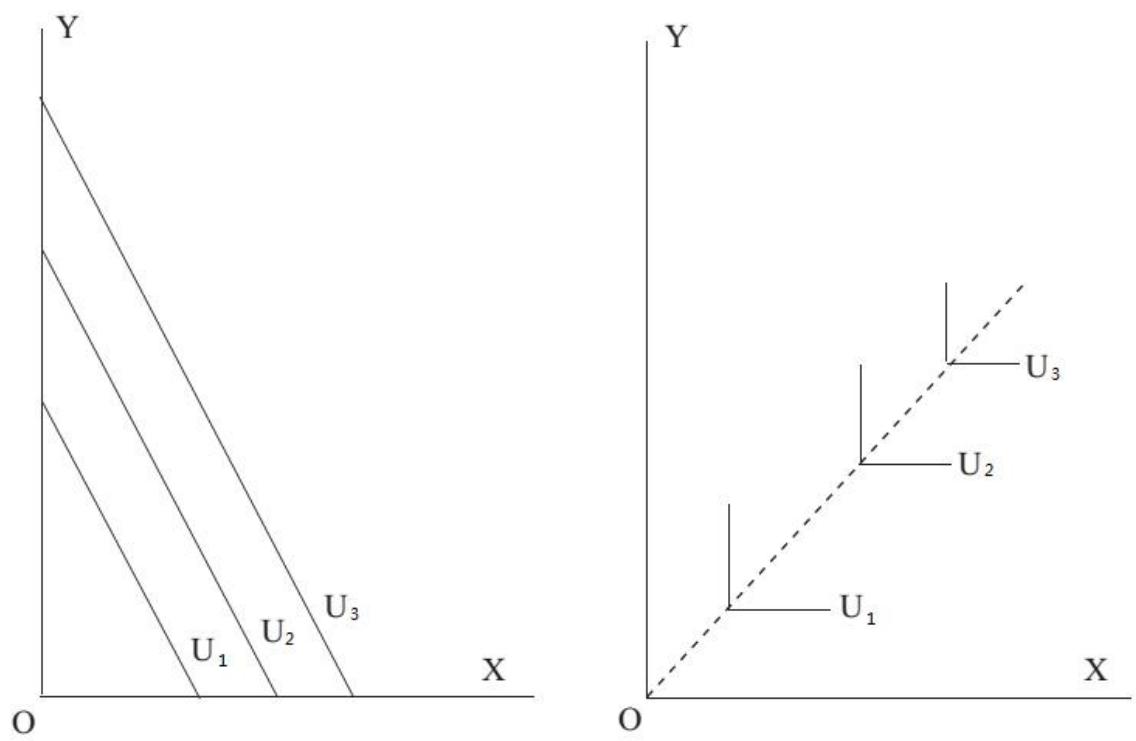


Figure 10-5

Section 4: The Optimum of the Consumer

In the above sections, only the postulates of consumer theory on human's intentions or preferences are discussed. In order to further find the optimum of the consumer or the consumption equilibrium, constraints must be included. Indifference curves only provide the following information: the consumer is indifferent to some points (baskets of goods) which are on the same indifference curve, and the further away from the origin, the higher the use value the indifference curve represents. Due to the human's nature of self-interest, if there are not any constraints, the consumer will choose countless points on the indifference curve that is infinitely far from the origin. Obviously, such an answer is meaningless and does not help to infer the consumer's behaviors.

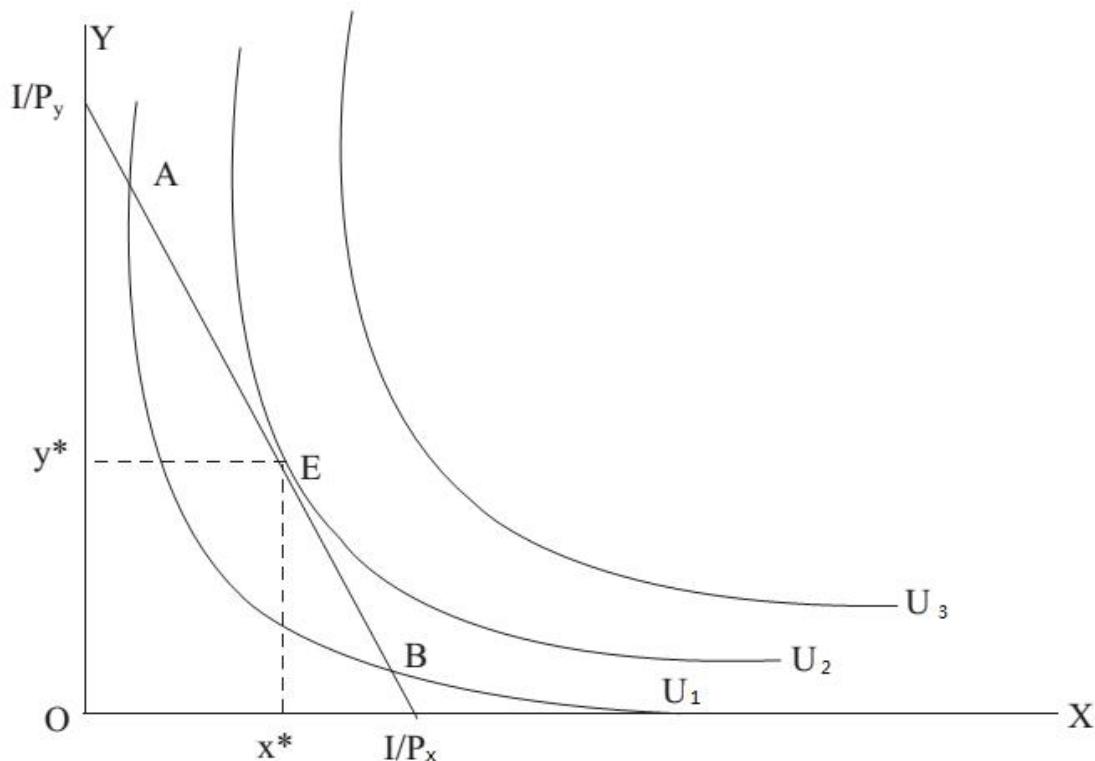


Figure 10-6

Next we will introduce the "Budget Line" which is used to reflect the constraints faced by the consumer. The straight line through the points A, E and B in Fig.10-6 is the budget line, which is drawn according to the equation of $P_xX + P_yY = I$. The economic implication of the budget line is that the consumer is faced with two constraints: income and price (the relative price of X and Y). If there were no constraint of income, the consumer could buy whatever he wants, so he would not need to choose. With the restrictions of "self-interest" and that "more is preferred to less", he could consume infinite quantities of both goods. Or even though his income is limited, but if the goods were free, the limited income would not be a constraint, and the consumer could also consume infinite quantities of both goods. Only when his income is limited and he must pay a price to buy the goods, need the consumer think about the following questions: how should I allocate my limited income to buy different goods for consumption, which is optimal to me? According to the analysis on preference in previous sections, the optimum of the consumer is to maximize the use value (utility), which is the implication of self-interest, under the constraints of income and price.

Let's look at the geometric solution first. As shown in Fig.10-6, the budget line and the indifference curves are drawn in the same rectangular coordinate. There are three kinds of relationships between the indifference curves and the budget line: intersection (e.g. U_1), tangency (e.g. U_2) and separation (e.g. U_3). Obviously, the indifference curve of U_3 is the farthest from the origin among the three curves, representing the highest use value and is most preferred by the consumer. However, it is a pity that it is not attainable for the consumer, because it is separated from the budget line, the economic implication of which is that every point on it cannot be affordable by this consumer under the constraints of his income and the current relative price of good X and Y. Everyone wants to live in a luxury villa, enjoy the delicacies and expensive food every day, and take a cruise to travel around the world, which is precisely represented by the indifference curves separated from the budget line such as U_3 that is too good to be affordable.

For the other two indifference curves, the arc from point A to point B of U_1 represents the baskets of goods that is affordable by the consumer, but this curve is the nearest to the origin among the three ones, which means that the use value it represents is also the lowest, so the consumer is not interested in it. The rest is the indifference curve of U_2 , and there is only one point on it that is affordable by the consumer can afford, which is point tangent to the budget line. One point is enough, and it is exactly the optimum of the consumer, because it is the only basket of goods that is affordable by the consumer on the farthest indifference curve from the origin that is attainable for the consumer under the current constraints.

Then let's turn to the corresponding algebraic solution. Since the geometric solution is the tangent point of the indifference curve that is tangent to the budget line, the algebraic solution is to find where the slopes of them are equal. The budget line is a straight line with the equation of $P_xX + P_yY = I$, so the slope of it is $\frac{P_x}{P_y}$, the relative price of the two goods. How about the slope of the indifference curve? According to calculus, the slope of a curve is the derivative of $\frac{dy}{dx}$, which is called the "marginal rate of substitution in consumption" (MRSc) in economics. The economic implication is that if the use values of the basket of these two goods keep unchanged, when the consumption of one good is increased, it is necessary to reduce the consumption of the other good, which means the substitution of the consumption of the other good with that of one good. Thus, $\frac{P_x}{P_y} = \frac{dy}{dx}$ is the algebraic solution to the optimum of the consumer.

The geometric and algebraic solutions mentioned above are both "interior solutions" in mathematics, but there are also "corner solutions". In Fig.10-7, the indifference curve of U_2 is tangent to the budget line in the negative domain, which obviously could not happen in reality. Therefore, the consumer can only choose U_1 as the second best choice, because it is the farthest one from the origin (and the closest one to U_2) among all the indifference curves that intersect the budget line in the positive domain. The intersection point of U_1 and the budget line is the same intersection point of the budget line and the vertical axis, the economic implication of which is that this consumer will spend all his income on good Y, and does not buy any good X at all. In reality, there are a large number of goods that a consumer does not buy at all, such as luxury villas, global tourism which are too expensive, as well as the goods which are very cheap with extremely poor quality, and the goods which are of no use to himself (such as a novel written in Malagasy). They are all the examples of good X in this Figure. By the way, it is necessary to have a broad understanding of good X and Y. For example, X in the corner solution can be regarded as all the goods the consumer does not buy, while Y can be regarded as all the other goods he will buy more or less. Thus, it is not unrealistic for this model of two goods.

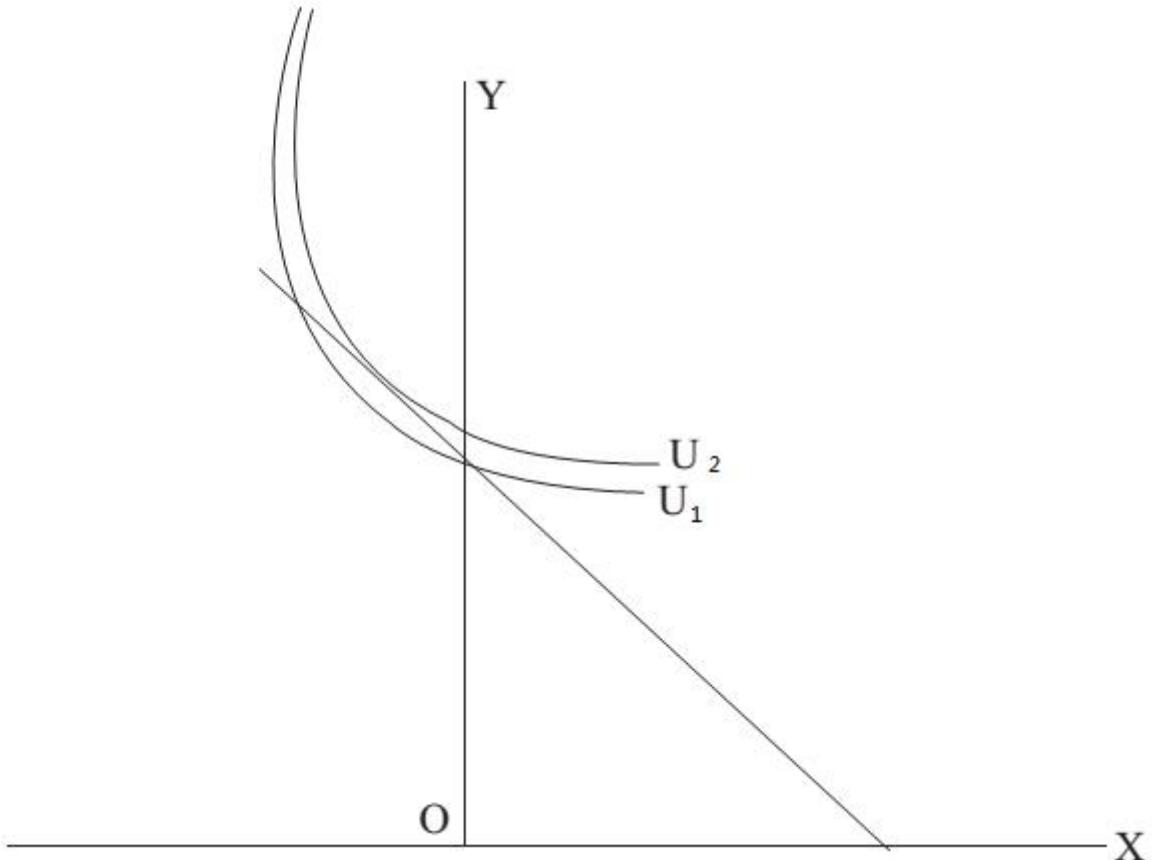


Figure 10-7

Attention: there is a constraint implied in the above analysis of the optimum of the consumer that there is no borrowing or lending in the world simulated by this model. In reality, people will not spend all their incomes now, which is usually not the optimum. They will choose between present consumption and future consumption, which will be analyzed by the interest theory in the later lecture. Now it is supposed that the present income must be spent completely at present, or else it will rot and lose its value in the future (imagine that this income is represented by fresh fruits and vegetables), so there will be no borrowing and lending in the word, which means you cannot lend the present income to others for his present consumption in exchange for his repayment to you in the future for your future consumption. And the current consumption cannot exceed the current income either, because you cannot borrow other's present income for your present consumption.

Section 5: The Changes of the Optimum of the Consumer

In the previous section, it has been explained how to find the optimum of the consumer: based on the two axioms about preference and the concept of utility (or use value), indifference curve is derived, the contingent point of which and the budget line reflecting the two main constraints of income and price determines the optimum of the consumer or the so-called consumption equilibrium. If the constraints change, the optimum or equilibrium will change too. This will be explained in this section.

Since the budget line reflects the two main constraints of income and price, the equilibrium changes caused by the changes in income and price are discussed separately below. Let's look at the income change first. As shown in Fig.10-8, if income rises from I to I' while prices are held constant, the budget line shifts outwards parallel to itself, from BL to $B'L'$. The new optimum position is at point E' , where the budget line L' is tangent to the higher indifference curve U_1 . As can be seen from Fig.10-8, points E , E' , E'' are equilibrium points when the income is I , I' , I'' respectively. Connecting all these equilibrium points, a curve called "Income Expansion Path (IEP)" can be drawn. Each point on this curve is an equilibrium point, only reflecting different incomes.

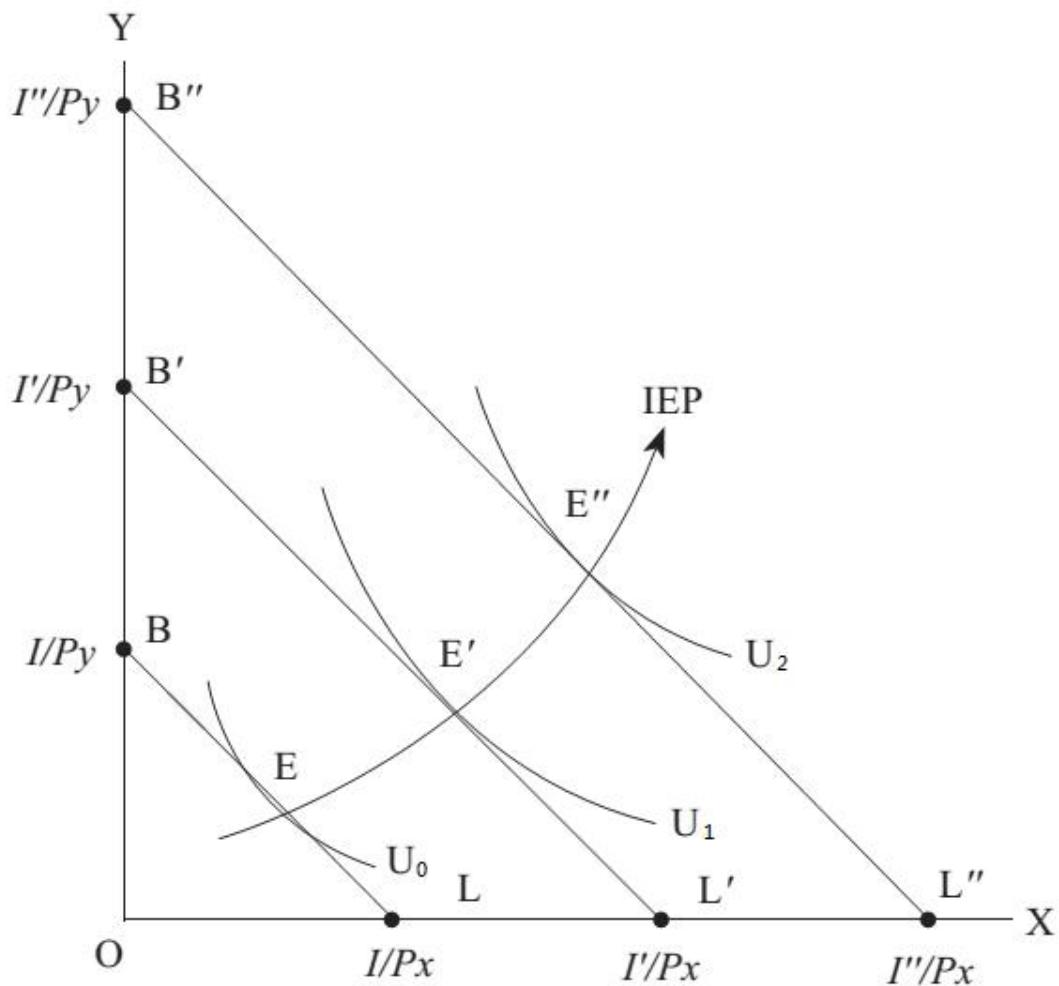


Figure 10-8

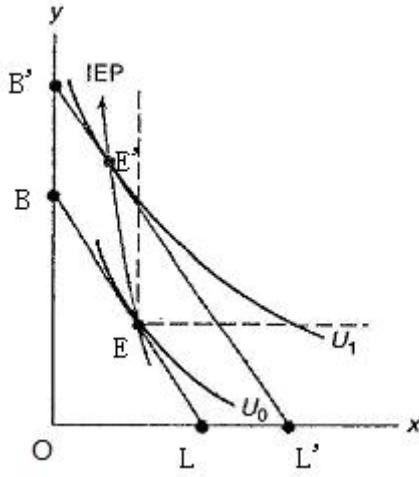


Figure 10-9

The shape of IEP can be that in Fig.10-8 or be that in Fig.10-9. For the IEP in Fig.10-8, when income increases, the consumption of both X and Y rises. However, for the IEP in Fig.10-9, when income increases, only the consumption of Y rises, while that of X falls. Both goods in Fig.10-8 and Y in Fig.10-9 are called superior goods, while X in Fig.10-9 is called inferior good. Caution: an inferior good is not a good of poor quality, but just refers to the good that the consumer will buy less when his income rises. For example, a bicycle may be a typical inferior good. One chooses to ride bicycle when his income is low, and change to car when his income rises, but his bicycle is not necessarily of poor quality. Now in many developed countries, when people's incomes are higher, they have become more environmentally friendly, and there is a tendency to return to bicycles. In this case, it is car that has become an inferior good!

The IEP shows how consumption baskets (combinations of goods) change as income rises or falls. The Engel Curve pictures the effect of income changes upon a single good X with the horizontal axis as income and the vertical axis as the quantity of X, such as Fig.10-10. The Engel curve is named after a German statistician Engel, because his research on the relationship between income and consumption is the most famous. Obviously, EC1 represents a superior good, while EC2 represents an inferior good.

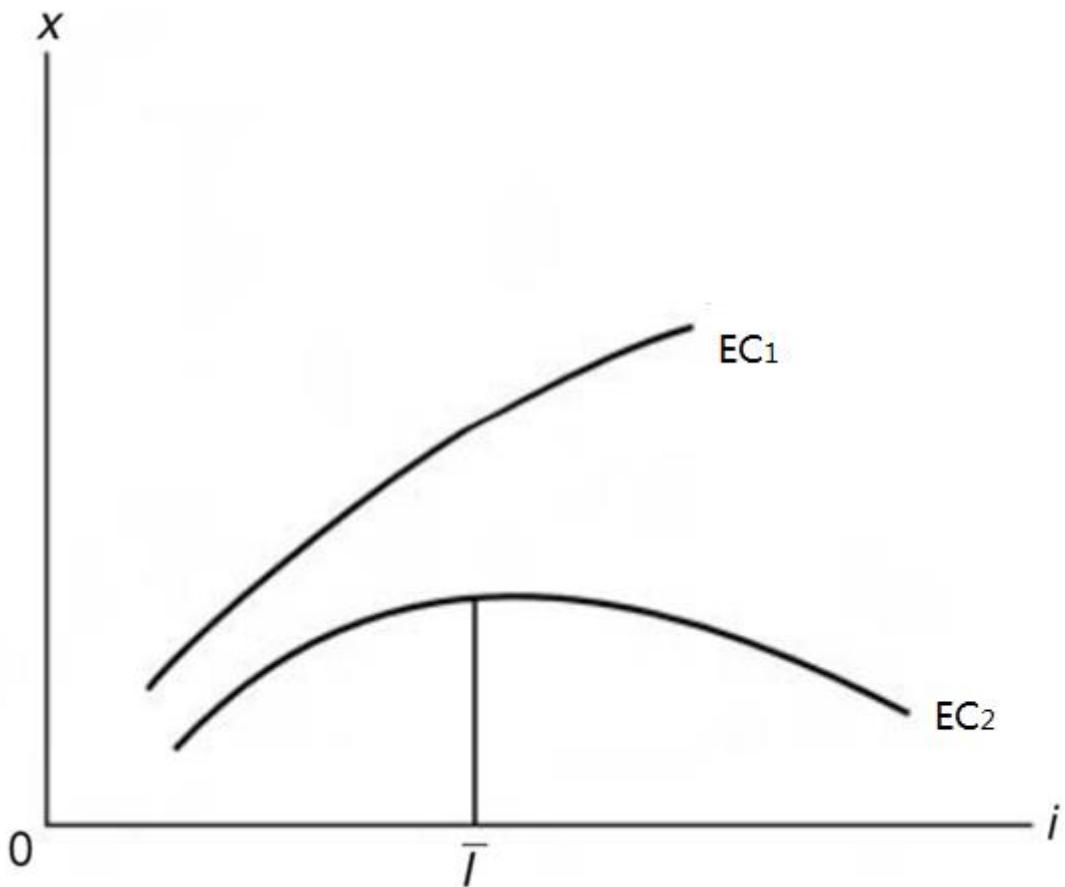


Figure 10-10

Next, let's look at the equilibrium change caused by the change in price. The slope of the budget line in Fig.10-11 represents the relative price of X and Y, so the change of it reflects the change in price, which will also cause the change of tangent point of the indifference curve. The points E, E', E'' in Fig.10-11 are equilibrium points with different prices. Connecting all of them, a curve called "Price Expansion Path (PEP)" can be drawn. Each point on this curve is an equilibrium point, only reflecting different (relative) prices. In Fig.10-11, only the (absolute) price of X changes, so the intersection point of the budget lines and the vertical axis has not changed, and just the intersection points of them and the horizontal axis have changed. The PEP drawn under such constraint (the absolute price of Y remains unchanged) starts from point B, and it is impossible to be higher than the vertical magnitude of point B (the horizontal dashed line in Fig.10-11). However, in reality, the (absolute) price of X and Y can both change at the same time, and then the intersections of the budget lines and the both axis will all change.

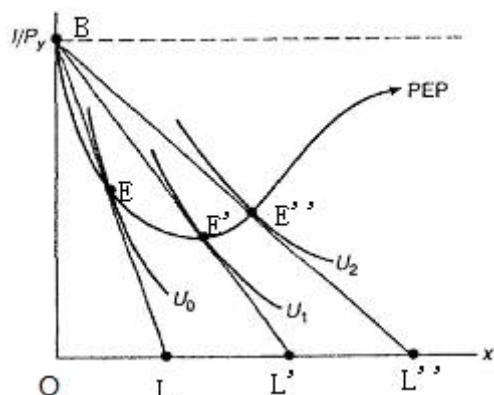


Figure 10-11

Draw both the IEP and PEP in a same coordinate system, and have a good understanding of this statement: every point in this coordinate plane must be on a certain IEP or PEP, which means that it is equilibrium. As a result, those who are used to thinking in mathematics (geometry) may be able to understand what is said in Lecture 7 that it is always equilibrium. If there is a case that seems to be disequilibrium, there must be some constraints that are neglected. Taking all the constraints into account, it is always equilibrium. Under different constraints, there are different equilibriums. As shown in Fig.10-8 and 10-11, point E' is disequilibrium with respect to the budget line of BL' (reflecting the constraint of a certain set of income and price), but it is equilibrium with respect to B'L' in Fig.10-8 or BL' in Fig.10-11 (reflecting the constraint of another set of income and price). If one thinks that the point E' is disequilibrium, he only misunderstand the relevant constraint.

The PEP shows how consumption baskets (combinations of goods) change as price rises or falls. The demand curve pictures the effect of price changes upon a single good X with the horizontal axis as the quantity of X and the vertical axis as price, such as Fig.8-1 in Lecture 8. MSE comes up with such a complicated consumer theory just to deduce the law of demand law as a theorem instead of an axiom.

Section 6: Are there Giffen Goods?

Just as for the Engel curve derived from IEP, when income rises, the quantity of good X will rise (superior good) or fall (inferior good); for the demand curve derived from PEP, when price rises, the quantity of good X will fall or rise. However, the latter is not in line with the law of demand!

Fig.10-12 illustrates how such a case that is not in line with the law of demand happen geometrically. Here is the economic implication. There are actually two kinds of effects for a price change: one is substitution effect, the other is income effect.

Substitution effect means when the price of a good rises, people will change to consume its substitutes. Under the constraint of constant income, all goods are substitutes, because if one wants to buy more of one good, he has to buy less of the other good to free up income.

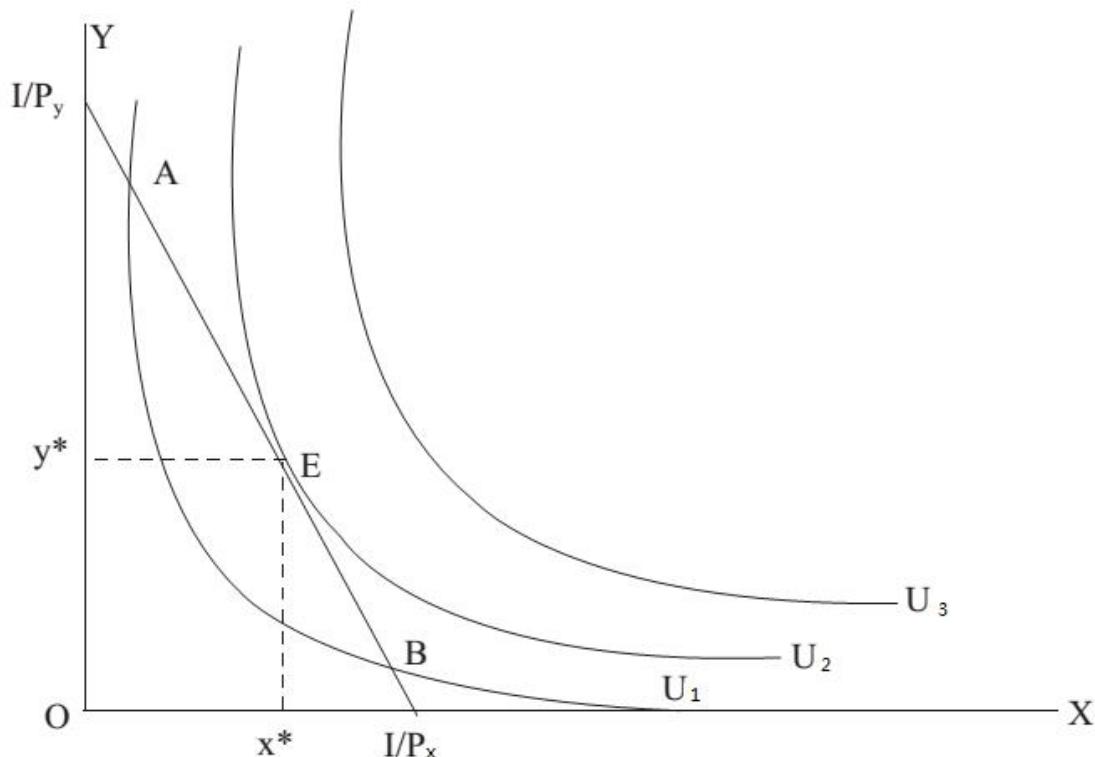


Figure 10-6[^1]

Income effect means when the price of a good rises, people's real income falls with other factors unchanged, which will cause people increase or decrease the consumption of the good. Suppose one's income is RMB1000, and the price of a bottle of water is originally RMB1. If he spends all his income on water, he can buy 1000 bottles of water. Then the price of a bottle of water rises to RMB2, he can only buy 500 bottles of water now, which is the same as if the price of a bottle water keeps RMB1 while his income falls to RMB500. Therefore, a rise in the price is similar to a fall in income. The purchasing power of monetary income is called real income in economics. For example, when the price of a bottle of water is RMB1, the real income of RMB1000 is 1000 bottles of water. So when the price of a bottle of water is RMB2, the real income of RMB1000 falls to 500 bottles of water. Thus, changes in the price of a good will cause changes in real income.

The substitution effect is based on the premise that "real income remains unchanged", then it can be separated from the income effect, and it is easy to understand why there would be cases that are not in line with the law of demand. The price and quantity in the substitution effect must change in the opposite direction, which is determined by the postulate of self-interest. However, the price and quantity in the income effect do not necessarily change in the same direction. In the

case of inferior goods, they change in the opposite direction. Thus, when the price of inferior goods rises, the substitution effect leads to a decrease in quantity. However, a rise in the price is equivalent to a fall in real income which causes the consumption of inferior goods to fall. As long as the income effect exceeds the substitution effect, there will be a rise in the net change of quantity.

This kind of goods that is not in line with the law of demand is called "Giffen Goods", because it is a British knight named Giffen who firstly proposed this possibility to Marshall when he revised his "Principles of Economics". Marshall wrote it down in his book and named it after him. In fact, Giffen had never seen such Giffen goods with his own eyes, but just proposed such a possibility in logic, and gave an imaginary example of Irish peasants' potatoes. Irish peasants were so poor that their income came mainly from the sale of potatoes and they also mainly fed on potatoes and a little meat. If there were a serious crop failure of potato in a year, the price of potato would rise dramatically, but the income of Irish peasants would drop dramatically. Although the price of potato rose, the production quantity of potato fell even more, so the income still dropped. These poor peasants could not afford even a little meat which was substituted by more potatoes, so the price of potato rose and it seemed to be not in line with the law of demand.

This imaginary example of Irish peasants' potatoes proposed by Giffen has mistakenly become a fact that he had really seen. The following reading material "Was Bread a Giffen Good?" is about the investigation by the economic historian to see if the bread of English rural laborer household which is similar to the potatoes of Irish peasants was a Giffen good. The result of it shows that bread was not a Giffen good. In fact, No economist has found any Giffen goods in reality so far.

READING MATERIAL: WAS BREAD A GIFFEN GOOD? ¹

Some historians have suggested that bread was a Giffen good for English rural laborers at the end of the eighteenth century.

A study by Roger Koenker casts doubt on this claim. ² In the late eighteenth century, the limited transportation network in Britain meant that prices differed considerably across communities. These differences, as well as changes over time, provided data for estimating the demand for bread.

The demand function of a typical English rural laborer household was estimated as follows:

$$\text{Quantity of bread} = 0.40 + 0.41 \text{ Family size} + 0.024 \text{ Weekly expenditures} - 0.35 \text{ Bread price} + 0.57 \text{ Meat price}$$

Here bread quantity was measured in loaves per week, bread price in pence per loaf, and meat price in pence per pound. Weekly spending (in pence per week) was used as the measure of overall income.

The negative coefficient (-0.35) for bread price shows that bread was not a Giffen good for these consumers. Indeed, the positive coefficient for weekly spending (income) suggests that both bread and meat were normal superior goods. (As the text showed, although an inferior good can be a Giffen good, a superior good cannot be.) The positive coefficient on the meat price means that an increase in the price of meat increased consumption of bread, suggesting that bread and meat were substitutes. Consumers, buying less meat in response to the higher meat price, bought more bread.

As mentioned in Lecture 8, there was a debate in China about whether the law of demand had been refuted. At that time, a large number of people claimed to have seen Giffen goods which were not at all if closely examined. As can be seen from the above discussion, Giffen goods are strictly defined in economics. The goods with the consumption that appears to rise along with the rise in price are not defined as Giffen goods. They must be inferior goods firstly, and then the

income effect must exceed the substitution effect. The higher the tuition (price) of an MBA degree is, the more people apply for it. The more expensive the luxury is, the more consumers buy it. The higher the stock price is, the more investors buy in ... These phenomena seem to refute the law of demand, but they cannot be regarded as Giffen cases. How likely are MBA degree, luxury, stock ... to be inferior goods which means people will consume less of them when they become richer? They cannot even satisfy the first requirement (inferior goods) of Giffen goods.

However, despite the above ignorant arguments, there are still following questions in economics: are there Giffen goods in reality? Has the law of demand been refuted by Giffen goods? There are three levels of answers to these questions.

The first level of answer: from the perspective of mathematics (geometry), there do logically be Giffen goods. Therefore, the law of demand cannot be deduced from consumer theory, unless it is supplemented with the supposition that there are no Giffen goods.

The second level of answer: from the perspective of reality, there cannot be Giffen goods. The reason is very simple. Although there are inferior goods in reality, it is almost impossible for the income effect to exceed the substitution effect. One needs to consume thousands of goods, so the expenditure of one good accounted for the proportion of his total income cannot be very high, and the income effect caused by the rise in the price of it is negligible, basically cannot exceed the substitution effect. Even for high-priced commodities such as real estate, they are bought through long-term savings at one time, so the price is actually not so high if decomposed into the proportion of the income in each month, which is much easier to understand when houses are commonly paid by installment nowadays. What is more, houses are seldom inferior goods, let alone Giffen goods.

Of course, it can be imagined that the extreme case such as the Irish peasants' potatoes proposed by Giffen would really happen in reality. The income effect of the price change in potato would be strong enough to overwhelm the substitution effect. However, as long as Irish peasants' income rose slightly, it were impossible to spend most of the income on one good. Moreover, the consumption of inferior goods would fall rapidly with the rise of income, so the impact of the income effect would also fall rapidly. Thus, there would be Giffen goods under so extreme circumstances that could only be imagined.

The third level of answer: from the perspective of market transactions in a society, there is impossible to be Giffen goods. If one would really buys more of a good when it is more expensive (the good is his Giffen good), others would raise the price of it to infinity and sell it to him. He would spend all his wealth on this good, and never sell it out again. In other words, Giffen goods will be eliminated out of the market by competition.

In conclusion, there are no Giffen goods in reality. Even there are, they are not important, and does not need to be cared about in economics. If it were not for MSE to deduce the law of demand with the more complicated consumer theory, but simply regard the law of demand as an axiom that needs no proof and only be accepted, the problem of Giffen goods would not be a problem at all!

1. This reading material is an excerpt from "Price Theory and Its Application", P118. [↩](#)

2. Roger Koenker, "Was Bread Giffen? The Demand for Food in England circa 1790," Review of Economics and Statistics, v.59 (1977). [↩](#)

Section 7: From Individual Demand to Market Demand

The demand curve mentioned before is the individual demand curve, while the market demand curve of a certain good shows the aggregate quantity demanded by all consumers together for it, and is obtained by summing the individual demand curves horizontally, as Fig.10-13 illustrates.

Note that Fig.10-13 is only applicable to illustrate the "Private Goods". If it is the market demand curve of a Public Good, it should be obtained by summing the individual demand curve vertically. The characteristics of public goods are non-exclusive and non-competitive in consumption, which means the same public good can be consumed by more than one person without interfering with each other.¹ For example, if one eats an apple, others are not able to eat it again, so apple is a "private good". By contrast, if one is listening to a song, others can also listen to it, so song is a "public good". Because the consumption of a public good is non-exclusive, the total demand of it cannot be obtained by summing all the individuals' demands, but the total income of it can be obtained by summing the prices paid by all individuals. For example, a song can be supplied to countless audiences, and the producer does not need to increase the quantities as more and more listen to it. By contrast, the producer of apple needs to increase the quantities supplied as more and more want to buy. Therefore, the income from this song is summation of the prices paid by all the audiences. That is why the market demand curve for public goods is obtained by summing the individual demand curves vertically, which means summation of the prices, instead of horizontally, which means summation of the quantities.

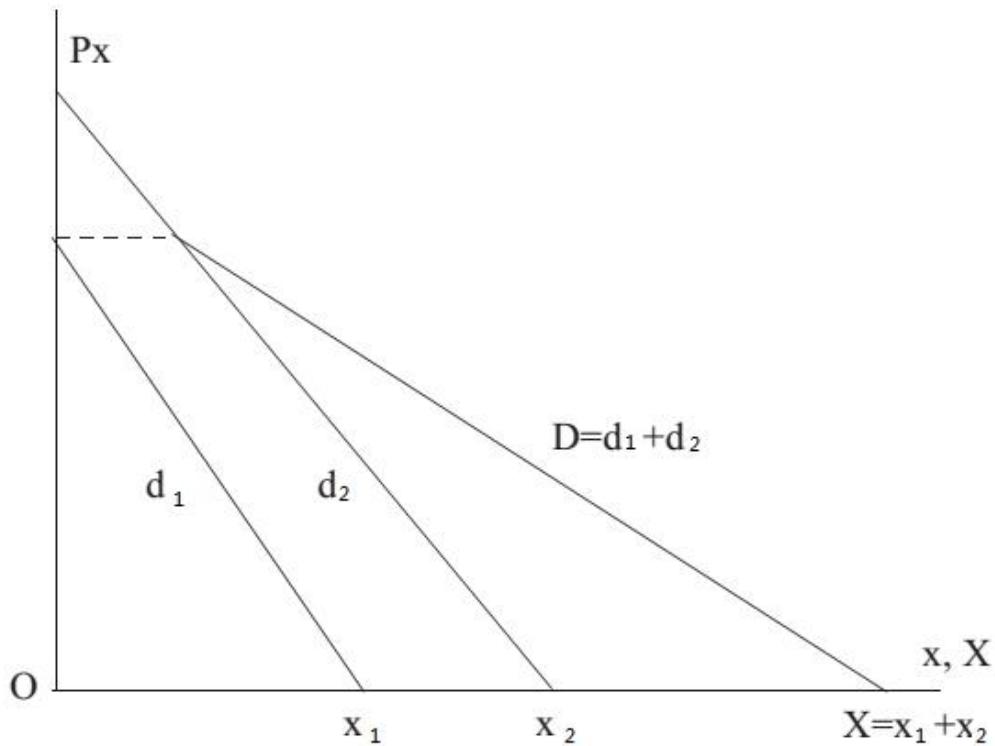


Figure 10-13

The non-exclusive consumption of public goods not only affects the market demand curve, but also makes it more difficult to protect the PPRs of public goods than private goods, which will be explained in detail later in this book.

1. Caution: do not confound public goods with common goods which are without PPRs or public utility supplied by government. Public goods can be owned and supplied by private individual. [e](#)

Section 8: Is There Theory of Elasticity?

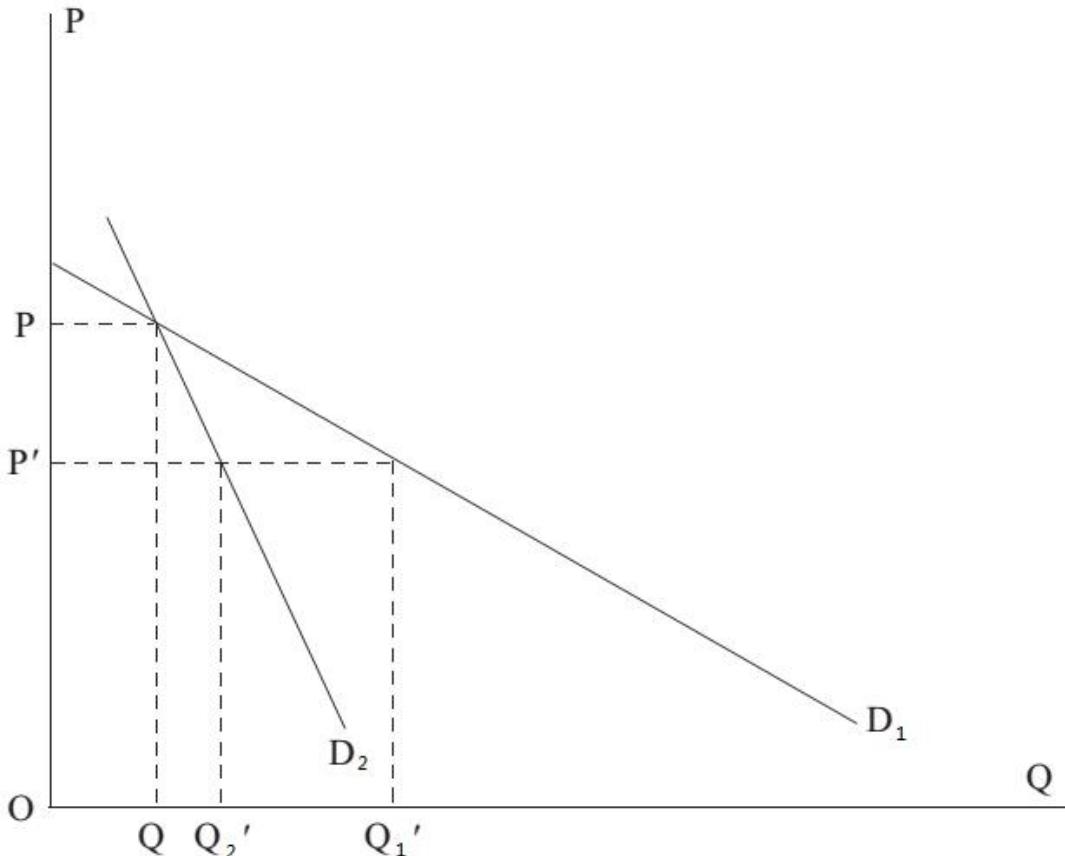


Figure 10-14

The law of demand only tells us that the demand curve has negative slope. As shown in Fig.10-14, both D₁ and D₂ are demand curves with different negative slopes, but the rises in quantity demanded caused by the same fall in price from P to P' are quite different for these two demand curves. Obviously, the rise in quantity demanded for D₁ with flatter negative slope is much more than that for D₂ with steeper negative slope. In reality, a slight change in the price of some goods does cause a larger change in quantity, while a large change in the price of other goods just causes a little change in quantity. In other words, there are different sensitivities to price changes for different shapes of demand curves, which are measured by "Elasticity" in economics.

The concept of elasticity in economics is again copied from physics. Elasticity in physics refers to the sensitivity of deformation of an object to changes in force. For example, if a spring will stretch a lot when pulled lightly, it is elastic. By contrast, if a spring will only stretch a little when pulled hard, it is inelastic. Human's skin is another typical example. It is very elastic because it will cave in when pressed lightly with a finger. By contrast, the surface of a table seems not change at all when pressed hard with a finger, unless it is watched carefully with a microscope, so the table is very inelastic.

It is Marshall again who copied the concept of elasticity from physics to economics, because the extent of quantity changes caused by price changes is equivalent to that of the deformation of an object caused by force. Elasticity coefficient is also used to measure elasticity, such as

$$\eta = \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} = \frac{\frac{\Delta Q}{\Delta P}}{\frac{Q}{P}} = \frac{\Delta Q}{\Delta P} \frac{P}{Q}.$$

$$\frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}}$$

is the definition formula of the price elasticity coefficient of demand. The numerator of it is the proportionate change in quantity purchased (i.e. the changing value divided by the initial value in quantity), while the denominator of it is the proportionate change in price (i.e. the change value of price divided by the initial value of price). This formula of elasticity measures a good's sensitivity to changes in price. The reason why an elasticity measure defined in terms of proportionate changes is used is to avoid the difficulty of being affected by the units of measurement. For example, if the direct measure of $\frac{\Delta Q}{\Delta P}$ is used, how to compare the price elasticity of cars in China and that of planes in US when there are different weight units and different currencies. However, the effects of different units can be eliminated by using proportionate changes, so the elasticity of any good in any currency can be directly compared.

Because the demand curve has negative slope, the price elasticity of demand must be negative. There are three kinds of elasticity based on the relationship between its absolute value and unity. If the elasticity measure exceeds unity in absolute value, a demand curve is said to be "elastic", which means the proportionate change in quantity is more than that in price. And the demand is said to be "inelastic" if the elasticity has an absolute value less than unity, which means the proportionate change in quantity is less than that in price. If the elasticity has an absolute value equivalent to unity, it is called "unitary" elasticity, which means the proportionate change in quantity is the same as that in price.

In addition, strictly speaking, the above formula of elasticity coefficient is used to calculate the arc elasticity. As the intervals shrink toward zero, the arc elasticity approaches the point elasticity, and the formula of elasticity coefficient becomes $\frac{dQ}{dP} \bullet \frac{P}{Q}$. The left part of this formula is the reciprocal of the slope of the demand curve, which shows that the elasticity and the slope of the demand curve are not the same. If the demand curve is a straight line, the slope on each point is the same, but the elasticity varies along the curve. So it is rarely advisable to describe an entire demand curve as elastic (or inelastic). Rather, one should say that demand is elastic (or inelastic) in the neighborhood of some given price-quantity point. However, at the same price level, the elasticity of demand curves with different slopes can be compared. As can be seen from the Fig.10-14, the flatter the demand curve is, the more elastic it is. On the contrary, the steeper the demand curve is, the more inelastic it is.

Why should we care about elasticity? Because if a good is elastic or inelastic will affect the pricing strategy of the producer. If a good is elastic, when the producer only slightly raises the price, the quantity will drop sharply. However, when the producer slightly decreases the price, the quantity will greatly rise. Because the revenue from the sale is the price multiplied by the quantity, it is better for the producer to adopt the strategy of cutting the price to promote the sales, which can bring about an increase in revenue.

By contrast, if a good is inelastic, when the producer raises the price substantially, the quantity will not drop much. However, when the producer decreases the price substantially, the quantity will not increase much. For the inelastic good, the strategy of price cutting will only decrease the revenue. There is a Chinese saying that "Low prices for grain hurt the peasants". The substantial rise in grain output during the harvest year will cause the price to drop sharply, but the consumptions will not increase much, thus causing the revenues of the peasants to fall instead of rise.

The concept of elasticity extends to many other fields. For example, from the supplements and substitutes discussed before, it can be seen that the price change in a good will not only affect the demand of itself, but also the demand for its supplements and substitutes. If the denominator of the coefficient formula (the proportionate change in price of a good) is changed to the proportionate change in price of another good, it is called the cross-elasticity of demand:

$$\eta_{xy} \equiv \frac{\frac{\Delta Q_x}{Q_x}}{\frac{\Delta P_y}{P_y}} \equiv \frac{\Delta Q_x}{\Delta P_y} \frac{P_y}{Q_x}$$

Unlike the price elasticity of demand which is always negative, the cross-elasticity of demand can be positive or negative. If good X is the substitute of good Y, the quantity of X will rise along with the rise in the price of Y, so the cross-elasticity of a substitute is positive. By contrast, if good X is the supplement of good Y, the quantity of X will fall along with the rise in the price of Y, so the cross-elasticity of a supplement is negative.

If the denominator of the coefficient formula (the proportionate change in price of a good) is changed to the proportionate change in income, it is called the income elasticity of demand:

$$\varepsilon \equiv \frac{\frac{\Delta Q}{Q}}{\frac{\Delta I}{I}} \equiv \frac{\Delta Q}{\frac{Q}{I}} \equiv \frac{\Delta Q}{\Delta I} \frac{I}{Q}$$

If the good is a superior good, the quantity of it will rise along with the rise in income, so the income elasticity of a superior good is positive. By contrast, if the good is an inferior good, the quantity of it will fall along with the rise in income, so the income elasticity of an inferior good is negative.

If the numerator of the coefficient formula (the proportionate change in quantity purchased) is changed to the proportionate change in quantity sold, it is called the price elasticity of supply:

$$\kappa \equiv \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} \equiv \frac{\Delta Q}{\Delta P} \frac{P}{Q}$$

, which seems the same as the formula for the price elasticity of demand

$$(\eta \equiv \frac{\frac{\Delta Q}{Q}}{\frac{\Delta P}{P}} \equiv \frac{\Delta Q}{\Delta P} \frac{P}{Q})$$

, but the Q in η means quantity purchased that is related to quantity demanded, while Q in κ means quantity sold that is related to quantity supplied. Since the supply curve is sloping upwards, the price elasticity of supply is always positive. There are also three kinds of price elasticity of supply based on the relationship between its value and unity. If the elasticity measure exceeds unity, a supply curve is said to be "elastic", which means the proportionate change in quantity is more than that in price. And the supply is said to be "inelastic" if the elasticity has a value less than unity, which means the proportionate change in quantity is less than that in price. If the elasticity has a value equivalent to unity, it is called "unitary" elasticity, which means the proportionate change in quantity is the same as that in price.

In fact, one can put any different data into the numerator and denominator of the elasticity coefficient formula to create various kinds of elasticity. For example, in international finance, if one wants to study how much the devaluation or appreciation of a country's currency affects the change in import or export volume, he can put the proportionate change in the value of currency value in the denominator, and the proportionate change in import or export volume in the numerator to calculate the monetary elasticity of import or export. Thus, there can be various kinds of elasticity in economics. Since the price elasticity of demand is the first elasticity, it is the default of elasticity if there is no specific statement.

However, elasticity is not really very useful in explaining phenomena or economic decisions, because so far economists have not found a reliable rule (theory) to determine how much the elasticity is in advance. If we cannot tell whether a good is elastic or inelastic until we observe the quantity drops significantly or slightly, we have not really used elasticity to explain the phenomena. The elasticity is again just a tautology, because it is always true ex post. The real explanation should be prediction ex ante. A useful theory of elasticity must be able to provide some clear principles, according to which one can know what are the constraints determining

whether a good is elastic or inelastic, and then examine the reality to predict in advance whether the good analyzed is elastic or inelastic.

It is generally believed in MSE that there are three kinds of determinants (constraints) that affect the price elasticity of demand.

(1) Availability of substitutes: Demand for a good will be more elastic the more numerous and the closer the available substitutes.

(2) Luxuries versus necessities: Demand for a “luxury” tends to be more elastic than demand for a “necessity”.

(3) High-priced versus low-priced goods: Along a linear demand curve, elasticity is higher at higher prices.

However, unfortunately, all these three constraints are not binding, and meaningless in inference of the elasticity. For example, the first constraint about the availability of substitutes means how many there are substitutes and how strong the substitutability is. As mentioned in the section about substitutes in Lecture 8, any two seemingly unrelated commodities can become substitutes under specific constraints. There is a real example in marketing: the game makers found that a chocolate popular in middle school students had become a competitor or substitute for their products. The pocket money of middle school students was basically a fixed amount, so when they bought more chocolate, they had to cut their spending on other goods including games.

The second constraint about the division between luxuries and necessities, but is it really so easy to distinguish luxuries and necessities? For a businessman, luxuries are actually necessities. He needs the luxuries (such as BMW) as a signal to convey the information that he is doing well in business, earning high income, and therefore is very capable. Thus here is again the problem: under specific constraints, any seemingly luxuries can become necessities.

As for the third constraint, the boundary of high-priced goods and low-priced goods is even more ambiguous. There is no absolute high price or absolute low price, which again depends on specific constraints.

A. A. Alchian (1914-2013) proposed a much clearer constraints than the above three constraints: time. With the passage of time, the information cost about substitutes for a good will decrease. As a result, the number of substitutes will increase, and the good changes from inelastic to elastic. Alchian called it “the second law of demand.”

However, Steven N. S. Cheung, his student, found that the fact of the undersea tunnel in Hong Kong had refuted this second law of demand. Every time the toll of the undersea tunnel in Hong Kong is increased, the traffic will always fall significantly at first, and then gradually rise to the previous level. It seems that the undersea tunnel changes from elastic to inelastic, which is precisely contrary to the prediction of the second law of demand. The reason is quite simple. As soon as the toll of undersea tunnel is increased, people turn to other transportation means (such as ferries). But after a period of time they will find that all other transportation means are far less convenient than the undersea tunnel, so they turn back to it gradually.

In other words, with the passage of time, the information cost about substitutes does decrease, but it does not logically always result in the rise in the number of substitutes for a good. It is also possible that people find there are no substitutes good enough when they are better informed. Thus, the number of substitutes for the good actually drops rather than rises.

In conclusion, until now there has not been a theory of elasticity that can tell a good is elastic or inelastic in advance by determining the relevant constraints. There has been a concept of elasticity which is useless in explaining or predicting phenomena.

```
},  
  svg: {  
    fontCache: 'global'  
  }  
};
```

Lecture 11: The Concept of Cost

From this lecture on, we will turn from consumer theory to theory of firm. As mentioned in Lecture 10, consumer theory of MSE is useless in explaining phenomenon, while there are full of serious mistakes in the theory of firm which will be discussed from this lecture on.

In many microeconomics textbooks nowadays, the issue about “the cause of firm” is included in the theory of firm. However, the cause of firm does not belong to the traditional theory of firm at all, but to the firm theory. “Theory of Firm” and “Firm Theory” does look very similar, but they are in fact rather different. Theory of firm studies the production process that how many inputs will produce how many outputs, focusing on the technical characteristics of production, and belongs to microeconomics (price theory). By contrast, firm theory studies the organizational form or contractual arrangement for production, focusing on the contractual or institutional characteristics of production, and belongs to new institutional economics.

In theory, production does not need to be carried out through firms. A street vendor or a store runner is an individual, not a firm. He also engages in the activities of supplying goods in the market. Of course, in modern society, a lot of production is carried out through firms rather than individuals, because it is more advantageous to organize individuals in cooperative production in the form of firms. However, the traditional theory of firm does not care about this phenomenon, and does not ask what makes firm as an organizational form more advantageous than the self-employed individual. It only studies how many production factors a producer (no matter he is an individual or a firm) inputs and then how many outputs he can get technically, and presents this technical feature mathematically with production function. Production function is just an equation reflecting the technical characteristics of production, so it is essentially not economics at all, but engineering. In reality, if different organizational forms or institutions are used along with the same technology, the same inputs will produce different outputs. It is precisely the problem that economics should study, because it cannot be answered by engineering.

The traditional theory of firm is essentially an engineering theory, not an economic theory. A theory studying production that truly belongs to economics cannot ignore the organizational form, contractual arrangement or institutional conditions of production, which is bound to fall within the scope of new institutional economics. In other words, the constraint of transaction cost must be considered and included in such a real economic theory.

However, in this and the following two lectures, we will first introduce the “pure” theory of firm that does not need to consider the constraint of transaction cost. It does not mean that there is no transaction cost in production, but that under specific circumstances, transaction cost is not the key constraint that determines the change in production and can be classified as “the other factors unchanged”. Later in this book, the concept of transaction cost will be introduced in detail, and then it will be also discussed in detail as a key constraint that determines the organizational form, contractual arrangement or institutional conditions of production. Only at that time can we analyze the cause of firm.

Section 1: Opportunity Cost

In this lecture, we will introduce one of the most important concepts in economics: cost.

Please think about the following questions: What is the cost of a student going to university? Firstly, he has to pay tuition. Anything else? Yes! In addition to tuition, he also has to give up the income from working.

The cost in economics or economic cost is different from that used in daily life which refers to accounting cost, historical cost or monetary expenditure. However, the economic cost refers to the highest income given up. That is why it is called “opportunity cost” which means that cost is the highest income among the opportunities given up.

Next, a numerical example will be used to illustrate the concept of opportunity cost. As shown in table 11-1 below, suppose there are two options A and B, and the income of option A is 100, while the income of option B is 200. What are the costs of these two options respectively? According to the concept of opportunity cost introduced above, the cost of A is the income of B which is 200, while the cost of B is the income of A which is 100.

	Income	Cost
A	100	200
B	200	100

Table 11-1

Suppose there is more option C with the income of 300 as shown in Table 11-2. What are the costs of A, B and C? If one chooses A, he gives up B and C, but which has the highest income? It is C, so the income of C instead of B is the cost, and the cost of A is 300. Similarly, the cost of B is also the income of C which is 300. What about the cost of C? If one chooses C by giving up A and B, the income of B is the highest among A and B, so the cost of C is 200.

	Income	Cost
A	100	300
B	200	300
C	300	200

Table 11-2

It is clear that the highest income given up is the cost. As long as the highest income remains unchanged, the cost will also remain unchanged no matter how the incomes of other options (including the income of the option chosen) change.

Caution: The cost of C is not the sum of the incomes of A and B, but the one with higher income. However, if one option can only be chosen by giving up multiple options, the incomes of them must be summed up. For example, what is the cost for a student to go to university? On the one hand, he has given up some money for paying tuition. Attention: tuition is cost, not because the student has spent money, but because he has given up the option of using the money to buy other goods. In other words, he has given up the use value of consuming other goods, and it is

also a kind of non-monetary income which is the cost of going to university. Cost must be seen from the perspective of giving up income, not from the perspective of expenditure, otherwise it will be confused with the concept of accounting cost or historical cost. On the other hand, the student has also given up the income of working. The student must give up these two options at the same time to go to university, so both incomes must be summed up. It is different from the numerical example in Table 11.2 where if A is chosen, B cannot be chosen, so only the higher income of B is the cost of C.

One does not need study economics to know that he should choose the option with the least cost. However, one needs to study economics to know that he should not choose according to (accounting) cost in daily life, but according to opportunity cost. For example, why are there more customers going to barbershop on weekends than on weekdays. The price of haircut is the same on weekdays as on weekends. If one chooses when to have a haircut according to the cost or expenditure in daily life, weekdays should be the same as weekends. The answer is: because people choose when to have a haircut according to the opportunity cost. If one goes to barbershop on weekdays, he not only gives up the money paying for the haircut, but also the income of working. By contrast, if he goes to barbershop on weekends, he does not need to give up the income of working. Therefore, the opportunity cost of haircut on weekends is actually lower than on weekdays. It is the law of demand that determines more people will choose to go to barbershop on weekends with lower cost! What is more, it is opportunity cost instead of only the price paid for a haircut that is the variable of price in the demand curve.

Section 2: Historical Cost is not Cost

There are some key points for opportunity cost.

Firstly, there is no cost if there is no alternative option. Because cost is the highest income of the option that has been given up, if there is no alternative option, there will be no income that has been given up and no cost at all.

Secondly, historical cost is not cost. The historical cost is the expenditure that has already been paid and has become history. Why is it not cost? The reason lies in the first key point above: there is no cost if there is no alternative option. Without time machine, human had not the option of going back to the past to choose once again. Therefore, the expenditure that has already paid is like the water that has poured out, and it cannot affect the current choice.¹

It is very important that historical cost is not cost, especially when making decisions in investment. Those who do not know or forget this key point will make mistakes. Suppose one has bought a stock at the price level of RMB100, and now the price falls to RMB80, should he sell the stock? He should not consider that historical cost of RMB100, but just consider whether the stock is expected to fall or rise in the future. If he expects the price of the stock will continue to fall, for example, to RMB50, the income and cost of his options is as follows:

	Income	Cost
A: Sell out	80	50
B: Not sell	50	80

Table 11-3

Next, a real example will be used to illustrate how to apply the concept of opportunity cost: SARS!

In the spring of 2003, SARS broke out in China. This disease spread by close droplets transmission, so wearing masks is considered to be the most effective preventive measure. As a result, the public rushed to buy mask. The price of mask soared, causing public criticism. The stores of mask sellers were believed to have been bought in before the outbreak of SARS, and the prices then were very low. The sellers who raised the price of mask were criticized as looting a burning house and even making a fortune on nation's trouble, and government was appealed to crack down on this so-called illegal price behavior.

Three economic questions can be made out on the price rise of mask during SARS:

(1) Even the price of mask has soared, but the demand for mask still dramatically increased. Has it refuted the law of demand?

(2) Suppose the price of mask before SARS was RMB1, and rose to RMB10 during SARS. Why did it happen? In other words, why did the suppliers sell out the masks at RMB10 instead of the stock price of RMB1? Cautious: do not ask whether mask suppliers are bad guys or whether they should raise the price of mask or not? It is economic question, not moral criticism.

(3) What would happen if government, influenced by public opinion, imposed price control prohibiting mask suppliers from raising the price?

Only the Question (2) will be explained here.

The masks bought before SARS were purchased at the price of RMB1 which was historical cost, not opportunity cost. What is the opportunity cost of the masks during SARS? Just with a look at the market of mask, anyone could see even if the price was raised to RMB10, the masks would be still sold out. Therefore, the options faced by a mask supplier were as follows:

	Income	Cost
A: 1RMB	1	10
B: 10RMB	10	1

Table 11-4

If the mask supplier continued to sell at the price of RMB1 as before SARS, his mask could be all sold out, and his income per mask would be RMB1. If he raised the price to RMB10, his mask could still be sold out due to the huge demand in the market during SARS, and his income per mask would be RMB10. According to the concept of opportunity cost, the cost of A is the income of B which is RMB10, while the cost of B cost is the income of A which is RMB1. It is easy to predict that the self-interested suppliers would choose B for cost minimization. Attention: In the absence of SARS, mask suppliers could not sell out the masks at RMB10, because there would not be so huge demand for the mask or there was so high use value of mask in preventing the transmission of SARS. Without SARS, the income of B is only 0, not 10, so the cost of A is also 0, and the self-interested suppliers will choose A instead of B for cost minimization.

It is obvious that opportunity cost can well explain the mask suppliers' behaviors no matter with SARS or without SARS. Especially during SARS, why would the mask suppliers set the price different from the purchase cost (historical cost) could not be explained by historical cost, but well explained by opportunity cost. In other words, the rise in price of mask during SARS has also proved eloquently that it is not the historical cost but the opportunity cost that determines human's behavior. It is precisely why in economics the concept of opportunity cost must take the place of the accounting cost or historical cost commonly used in daily life. It is not because economists like to be different, but because historical cost cannot explain the phenomena in reality, while opportunity cost can.

1. Some economics textbooks will also introduce a concept called "Sunk Cost" and carefully distinguish it from historical cost. However, they are in fact the same, and neither is (opportunity) cost. ↵

Section3: The Disturbance of Historical Cost

Some may ask: since historical cost does not affect human's choices, why is the concept of cost commonly used in daily life historical cost? Because under the constraint of information cost, there is still function of historical cost that human often use historical cost to estimate opportunity cost. As mentioned earlier, opportunity cost is not the actual expenditure, but the highest income given up. Since opportunity cost is the highest income given up, it is difficult to know exactly what the income is for the option that has not been chosen. Therefore opportunity cost is not easy to estimate, unlike historical cost which is what has happened and the value of it is fixed. In other words, generally the information cost of opportunity cost is higher than that of historical cost, so human tend to estimate opportunity cost based on historical cost, which is based on the assumption that the constraints that determined what had happened in the past will continue to determine what will happen in the future. As mentioned in Lecture 8, there is a school called "technical analysis" in the stock market that uses the path of the stock prices in the past to guess that in the future, which is precisely because the information cost for predicting the stock price is too high. It is the same as the explanation for human using historical cost to estimate opportunity cost. Just as it is wrong for the school of technical analysis" to assume that what had happened in the past will continue to happen in the future, it is also wrong to estimate the opportunity cost by historical cost. Human are not unaware of the problem with this assumption, but under the constraint of the high information cost, they can only choose to inaccurately estimate opportunity cost based on such a problematic assumption. As has been mentioned in Lecture 3, human are rational, which does not mean that they will not make wrong choices.

Steven. N. S. Cheung has given a real example of selling used camera lens, which can well illustrate how people are misled by historical cost and mistakenly estimate opportunity cost under the constraint of information cost. Cheung intended to sell a used camera lens in the United States in 1970s. He had bought the lens for 500 dollar a few years ago, and he thought the depreciation of it should be 200 dollar, so he offered it for 300 dollar in the newspaper. There were so many buyers competing to buy it that it was sold out for 400 dollar finally. Only later did he know that the new product of the lens was selling for more than 1,000 dollar in the market. He sold it too cheaply.

The purchase price of 500 dollar after deducting depreciation is 300 dollar, which is historical cost. The opportunity cost of the used camera lens should be the current price of new product after deducting depreciation, which is more than 800 dollar. Cheung made a wrong choice because he did not know the information of the current price of new product. Of course, he could investigate the current price of the new product before selling the used lens, and then he would not be misled by historical cost, which would take time and effort that is precisely information cost. Generally speaking, due to the technical depreciation caused by technological progress, even if there is no physical depreciation caused by the use, the price of a good usually tends to drop over time. However, the constraint of technological progress is not always the key one that determines the price; otherwise there would be no antiques which are the older the more valuable. If one saves the information cost by not investigating the current price, he will take the risk that the constraint of technological progress may not be the key one that determines the price. In the later lecture of "Information Cost", it will be pointed out that risk can be measured by information cost. If there are risks, it actually means that there are information costs. When one saves the information cost of investigation, he has to take the risk of making wrong decision which is also

another kind of information cost. When Cheung sold his used camera lens, he must have thought the information cost of the latter was lower, so he chose to take the risk.

His choice can be simulated by the table 11-5. Suppose there is not depreciation of the used camera lens for simplicity, there are two options: A is to investigate the current price, and B is to estimate the price directly based on the historical cost of 500 dollar. For option A, before the investigation, he estimated that there would be 1% of possibility that the lens had appreciated over the years, while there would be 99% of possibility that the price had not changed. And it would take time and effort to investigate the current price, which was assumed to be equivalent to 10 dollar for him. Thus, the income of option A is $1000 \times 1\% + 500 \times 99\% - 10 = 495$ dollar. For option B, there was no cost in investigating the current price, so the income is simply 500 dollar. Those who would minimize cost naturally chose B under such constraint. Attention: the market price of 1000 dollar known afterwards could not be regarded as the income of option A, because it was not known at the time of making decision. If it were known, there would be no risk or information cost about knowing the current price, and there would be no need to calculate the income of A with probability or mathematical expectation.

	Income	Cost
A: Investigate the current price	495	500
B: Estimate by historical cost	500	495

Table 11-5

Of course, if the second-hand commodity is something with high possibility of appreciation (such as antiques), the probability in the mathematical expectation will need to be adjusted, and option A is entirely possible to be the optimum with lower cost. Suppose the possibility a good appreciating to 1000 dollar is 50% instead of 1%, so the possibility of that the price of 500 dollar has not changed is 50%. Thus the income of investigating the current price becomes $1000 \times 50\% + 500 \times 50\% - 10 = 740$ dollar, as shown in the table 11-6.

Obviously, under the above constraint, people will definitely choose to investigate the current price, even if it takes time and energy for the investigation which is equivalent to 10 dollar, and will not be satisfied with estimating the price based on historical cost.

	Income	Cost
A: Investigate the current price	740	500
B: Estimate by historical cost	500	740

Table 11-6

In brief, as long as the information cost is low enough compared to the incremental income brought by the investigation, people will not choose to estimate opportunity cost with historical cost. And then it is clear that the essence of choice is based on opportunity cost rather than historical cost.

Looking back on the example of the masks during SARS: When SARS broke out, there were so many reports from the media that masks could prevent the droplet transmission, that the mask suppliers could easily know that even if they raised the price to RMB10, they could sell out the masks, which means the information cost about the opportunity cost of masks had fallen to 0. No wonder the mask suppliers no longer estimated opportunity cost with historical cost as before, and chose to raise the price directly based on opportunity cost.

Section 4: The Laboratory in Economics

From the discussion about opportunity cost and historical cost, readers should have a good understanding how to test an economic theory with facts. Under normal circumstances, it seems that suppliers' pricing behaviors are based on historical cost, because they carefully record the historical cost of various commodities in accounting books. In fact, they only estimate opportunity cost by historical cost under the constraint of high information cost. However, how do economists know that suppliers' pricing behaviors are actually based on opportunity costs rather than historical costs? Under normal circumstances, due to the high information cost, suppliers' behaviors of using historical cost to estimate opportunity cost cause historical cost cannot be separated from opportunity cost. Therefore, economists have to find some very special circumstances where the constraint of information cost becomes insignificant, so suppliers do not need to estimate opportunity cost by historical cost. Only then can historical cost and opportunity cost be clearly separated, and it is clear that it is opportunity cost that actually determines human's behaviors or decision-makings.

It is logically similar that before Galileo, it seemed that the heavier a body was, the faster it fell down, which meant the speed of a falling body was determined by constraint of its weight or mass. However, it is only because the air friction is an important constraint that affects the speed of the falling body in the environment of the earth. The air friction is related to the mass, or strictly speaking the density, of a body, so it interferes with the effect of gravity, which has seriously misled scientists for a long time.

In 1969, when man landed on the moon for the first time, the astronaut has done a physical experiment that he let go of an iron ball and a feather at the same height and at the same time on the surface of the moon. These two bodies were observed to land on the ground at the same time. In such an environment as the moon close to vacuum, the air friction is reduced to basically 0, and it is successfully separated from the effect of gravity. And then it is clear that it is gravity rather than the weight or the mass of a body that actually determines the falling speed of a body.

Economics is a science, which is essentially the same as natural science like physics. There are many constraints that affect the research objects. Scientists must try to separate different constraints to study their effects. The advantage of natural science like physics is that it can artificially create an experimental environment that separates different constraints in the laboratory, while economics, as a social science, cannot do so. Although there is so-called "experimental economics", this attempt to imitate natural science and conduct controlled experiments is fundamentally out of line with the nature of economics as a social science. The constraint in economics is social, not physical, and the nature of sociality will be lost if it is separated from the society in reality. However, economics has its own laboratory which is the reality. Economists do not need any controlled laboratory. What they need is to open their eyes and look for those special circumstances in reality which can meet the requirement of separating the different constraints. For example, SARS is such a special circumstance when the information cost of mask suppliers' pricing could be reduced to 0, meeting the requirement of separating historical cost from opportunity cost.

When Galileo discovered the law of free falling body, the physicists did not have much better experimental conditions than the economists nowadays. They had not the advanced technology that later sent the astronaut to the moon to drop an iron ball and a feather at the same time. Neither had they a laboratory to conduct experiments with vacuum environment. Galileo's experiment was to plane a piece of wood very smooth and set it up as all kinds of slopes with different leans, and let a ball roll down from the slope. He used this experiment to replace the

experiment of free falling body, accurately calculating the relationship between distance and time, and making a decisive experimental conclusion.

As for economists, they should never expect to have a laboratory that can really stimulate the reality including all the key social constraints. They should try to know more about the reality, looking for the extreme cases that can meet the specific requirements.

Do not worry there are too few such extreme cases. It seems that such cases as SARS are very rare to happen. However the cases when the information cost falls to zero, separating historical cost from opportunity cost do not just the masks during SARS. For example, antique is also such a case!

Antique is a good that most typically implies that historical cost has no effect on pricing. The historical cost of a bowl made in the Tang Dynasty is negligible compared with the price of it in the market nowadays. Anyone will realize that historical cost has nothing to do with the price of an antique. There has been a long and difficult process for economists to understand opportunity cost. The classical economists, just like Aristotle in ancient Greece, were misled by the superficial phenomena to think that historical cost was the determinant of the value or price of a good, so the Theory of Labor Value was rather popular in classical period. There are two mistakes in this theory: one is that labor is regarded as the only source of the value of a good, and the other is that historical cost (the labor input during the production) is cost.

The former mistake will be discussed in the next lecture. Here the latter mistake will be discussed. Classical economists felt very confused with antiques, because they were obvious to refute the theory of labor value which thought that the value of a good should be determined by the labor input during the production. Ricardo had to divide all goods into two kinds: ordinary goods and antiques, arguing that the theory of labor value theory could apply to ordinary goods, while the value of antiques should be explained by another theory. However, Ricardo's contraction the application of the theory of labor value could not save it from being refuted by such goods as the masks during SARS that are not antiques. What is more, the concept of opportunity cost can explain the pricing of all good, no matter they are antiques or the masks during SARS, so it has stronger explanatory power and more widely application than the theory of labor value which essentially explains the pricing of goods with historical cost. That is one of the reasons why the theory of labor value has been abandoned.

Lecture 12: The Concept of Rent

Section 1: From Income and Cost to Rent

In the previous lecture, when explaining the concept of cost, there was a numerical example which is shown here in Table 12-1 with an additional column of "rent" calculated by subtracting (opportunity) cost from income.

	Income	Cost	Rent
A	100	300	-200
B	200	300	-100
C	300	200	100

Table 12-1

Rent of the third column in Table 12-1 is often called "economic profit" by MSE and distinguished from accounting profit which is calculated by subtracting accounting cost from income, just as opportunity cost introduced in the previous lecture distinguished from accounting cost (historical cost). The concept of cost in economics must refer to opportunity cost, so the word "opportunity" need not be mentioned in economics. Similarly, the concept of profit in economics must refer to economic profit, so the word "economic" does not need to be mentioned in economics. However, it is not very accurate for MSE to use "profit" to refer to the difference between income and cost. It is much more accurate to use "rent". And the term of "profit" should be used to refer to another concept different from rent or profit in MSE, which will be discussed in the last section of this lecture.

Rent of the third column in Table 12-1 is often called "economic profit" by MSE and distinguished from accounting profit which is calculated by subtracting accounting cost from income, just as opportunity cost introduced in the previous lecture distinguished from accounting cost (historical cost). The concept of cost in economics must refer to opportunity cost, so the word "opportunity" need not be mentioned in economics. Similarly, the concept of profit in economics must refer to economic profit, so the word "economic" does not need to be mentioned in economics. However, it is not very accurate for MSE to use "profit" to refer to the difference between income and cost. It is much more accurate to use "rent". And the term of "profit" should be used to refer to another concept different from rent or profit in MSE, which will be discussed in the last section of this lecture.

Look back at the numerical example in Table 12-1. Firstly, for the income of the first column, which option will be chosen? Obviously, C will be chosen because it is the option with highest income. Secondly, for the cost of the second column, which option will be chosen? Obviously, again C will be chosen because it is the option with least cost. Finally, for the rent of the third column, which option will be chosen? Obviously, C will be still chosen because it is the option with highest rent. In other words, income maximization, cost minimization and rent maximization are all the same.

Under different circumstances, sometimes the information about the income of various options is easily available, while sometimes it is the information about cost, or sometimes it is that about rent. Therefore, these three different angles can be respectively adopted according to specific circumstances. In addition, different economists may be good at analysis from different angles. For example, Ronald Harry Coase (1910-2013) is especially good at analysis from the perspective of cost, while Steven Cheung likes the perspective of rent. As will be explained in the later lecture

on the "Interest Theory", wealth is the sum of the discounted values of future income, so under a certain discount rate (interest rate), income maximization is same as wealth maximization. In other words, with interest taken into account, wealth maximization or income maximization is the most common angle to analyze different options for investment.

Section 2: The Changeable Rent

Rent is amazing in that it seems sometimes large and sometimes small and sometimes even zero!

Take the numerical example in previous section as an example again, but suppose the options are about how to use a piece of land, as the Table 12-2 shows.

	Income	Cost	Rent
A: Agricultural Land	100	300	-200
B: Industrial Land	200	300	-100
C: Residential Land	300	200	100

Table 12-2

Option A is to use the land to grow crops, and the income is 100. And option B is to use the land to build factories, and the income is 200. And option C is to use the land to build houses, and the income is 300. The respective cost and rent of these three options are also the same as those in Table 12-1 of previous section.

However, such division of options is rather rough. Option A is to use the land to grow crops, but what on earth is grown, wheat or rice? If it is rice, what is precisely the variety? Is fertilizer used? If yes, how much is used? Is one-harvest or two-harvest per year? In other words, there are actually countless sub-options in option A. It is the same for option B. Option B is to use the land to build a factory, but what kind of factory on earth is built? What is the plot ratio? Is it a bungalow or a building? How many floors? Is the factory of textiles, toys, steel, cars or software programs? It is the same for option C. Option C is to use the land to build a house, but what kind of house on earth is built? Is it a villa or an apartment? What is the plot ratio? How many floors? Is the building with or without elevators? How many households are there per floor? Is a household with one-hall-two-room or two-hall-three-room? Is the house with or without decoration?

As the option is subdivided, the rent will change. Suppose the option C is divided into two sub-options: villa and apartment, the above numerical example will become like this:

	Income	Cost	Rent
A: Agricultural Land	100	300	-200
B: Industrial Land	200	300	-100
C1: Residential Land-Villa	250	300	-50
C2: Residential Land-Apartment	300	200	50

Table 12-3

As shown in Table 12-3, the rent of the optimum falls to 50 from the original 100 simply because the option C is subdivided. Attention: The optimum of C2 here is exactly the same as the previous optimum of C. The option C previously describes building an apartment on the land roughly as "residential land", while the option C2 now more specifically describes it as "apartment". And as the description becomes more specific, it can be seen that there is the option C1 of villa between

B and C2. Thus, if the option C is subdivided, the suboptimum is no longer option B, but C1, so now it is the income of C1 instead of B becomes the cost of the optimum of C2. Since the income of C1 is higher than that of B, the cost of the optimum of C2 rises and the rent falls.

If the option C is further subdivided, until there are countless options within it, and then they are sorted by their incomes, the above numerical examples will become like this:

	Income	Cost	Rent
A: Agricultural Land	100	300	-200
B: Industrial Land	200	300	-100
C ₁ : Residential Land-Villa	250	300	-50
.....			
C _{n-1}	299.99999...	300	-0.0000...1
C _n	300	299.99999...	0.0000...1

(Note: n is infinity)

Table 12-4

The optimum of C_n can be described in detail as a 10-storied apartment with elevators, two households per floor, each household with two-hall-three-room, with refined decoration ... And Then a suboptimum of C_{n-1} can be always found, which is almost exactly the same as C_n except for a slight difference (such as a slight difference in decoration or in orientation), so the income of it is also only a little difference (299.99999 ..., with an infinite number of 9) from that of C_n's (300). With the option of C subdivided infinitely, the income of C_{n-1} is the cost of C_n, which is also infinitely close to the income of C_n, and the rent is infinitely close to 0 from the direction of the positive number axis. By contrast, the rent of C_{n-1} is infinitely close to 0 from the direction of the negative number axis.

There is another opposite direction for dealing with the options of how to use the land. Instead of infinite subdivision, the most rough division is to divide it into only two options like this:

	Income	Cost	Rent
A: Use	0	300	-300
B: Not Use	300	0	300

Table 12-5

If the land is left unused, the income will naturally be 0, so the cost of using it (specifically, building a 10-storied apartment on the land) will be also 0, and all the income will be rent.

Look back at the statement at the very beginning of this section: Rent is amazing in that it seems sometimes large and sometimes small and sometimes even zero! The option is in fact always the same one which is objective and has never changed. What has changed is how to observe the option and to what degree the option is subdivided.

It is logically similar that if observed from the galaxy, the earth is a particle which means it is a body only with mass and without volume. However, if observed much more closely, like from the solar system, the earth can no longer be regarded as a particle without volume, but as a blue planet. Then if observed within the atmosphere of the earth, it is a combination of mud, rock, water and air. At last, if observed the most closely with a microscope, the earth is a compound of hydrogen, oxygen, carbon and other atoms. It is always the same earth which is objective has never changed. What has changed is how closely to observe it.

Why are there different observations about the options? Because different problems need to be observed differently, just as different physics problems involving the earth need to be observed from different distances. For example, for the government, when it plans the use of a piece of land, it only needs to roughly divide the options into agricultural land, industrial land and residential land, and choose the option that can bring the highest rent to the society. If the land is planned to be used for building houses, the government will auction the land for residential use. The developers who come to bid will choose the option of villa or apartment depending on which can bring the highest rent. When the developers have bought the land and start construction, they will subdivide the options most specifically. Suppose they choose to build an apartment, they need to choose among all these options: what is the plot ratio? How many floors? Is the apartment with or without elevators? How many households are there per floor? Is a household with one-hall-two-room or two-hall-three-room? Is the house with or without decoration? It is subject to which can bring the highest rent.

As mentioned before, self-interested human will choose the option with maximal rent. It can be seen from the numerical example that although the rent varies from large to small or even to zero along with the options being more and more sub-divided, no matter how the rent changes, the optimum of rent maximization is always the same one, and the relevant rent is always non-negative. In fact, only the rent of the optimum is non-negative. In other words, people do not need to calculate the exact value of the rent, but only need to know whether it is negative. If the rent of an option is negative, it must not be the optimum. It is also why rent maximization (non-negative rent) is often a better angle than income maximization or cost minimization. In reality, there is constraint of information cost, and the formation cost of judging whether the rent is non-negative is usually the least.

Let us review the two extreme situations: one is the most infinite subdivision of options, and the other is the roughest division that there are only two options of "use" and "not use".

In the former case, the rent of the optimum is 0 and the cost is equal to the income. In other words, income is essentially cost. As mentioned in the previous lecture, the (opportunity) cost in economics is not expenditure used in daily life, but income which is the income of the sub-optimum given up rather than that of the optimum. However, it can be concluded from the above analysis that when the options are infinitely subdivided, the cost of the optimum is the income of itself.

In the latter case, the cost is 0 and the rent is equal to the income. In other words, income is essentially rent.

Deduced from these two extremes, when it is normal situation, rent is essentially cost. All income is cost, only part of it is the cost of the options observed, and part of it is the cost of the options not observed. In order to distinguish the options observed and not observed, the former is directly called "cost" and the latter is called "rent", but in essence they are both costs. Thus, all income is cost, and there is not such income as is not cost. It is a very important statement, because it is determined by self-interest and competition. And the implication of it is the "Zero-Profit Theorem" that will be introduced later.

Section 3: The Concept of Rent from the First Perspective

The concept of rent has not been introduced until now, because it is much easier to understand based on the explanation in previous section. There are two perspectives for the understanding of the concept of rent, which are actually the same.

From the first perspective of the historical evolution of the concept, rent is the income of production factor or the value of asset. Maybe many readers have guessed from the term of rent that it originally referred to the rent of land. That is right! In the period of classical economics, although the industrial revolution had begun in Britain, economists were still mainly concerned about the activities in agriculture. And land is the most important production factor in agriculture, so rent of land, as the income of land, was analyzed in detail by classical economists.

As mentioned in the previous lecture, classical economists advocated the theory of labor value which means the value of a good only comes from labor, so the income from a good should only be distributed to labor. However, when classical economists observed agricultural production, they faced a huge puzzle: why can a landlord collect rent? The rent of land is, of course, the income of land, but the landlord does not work by himself, and just leases the land to the farmers who work.

This obvious fact clearly has refuted the theory of labor value and hit classical economists hard. When a theory is refuted by facts, what a scientist should do is either to revise the theory or completely abandon it. In the previous lecture, we have introduced how Ricardo tried to revise the theory of labor value by narrowing the scope of application to exclude antiques which obviously refuted the theory regarding historical cost as cost. Of course, his revision failed to save the theory, because there are many goods other than antiques such as masks during SARS that also refute it. That is why economists completely abandoned the theory of labor at last, replacing historical cost with opportunity cost.

However, there is also another unscientific way: to use subjective value judgment to deny the objective facts. For example, when the price of mask during SARS rose sharply, people including laymen and incompetent or impudent economists would like to accuse mask suppliers of making profit when the nation was in trouble. They simply labeled the suppliers as immoral without any scientific explanation for the rise in price, which was essentially ostrichism refusing to face the reality. If this kind of ostrich policy became the basis for government's decision, it would be impossible to effectively solve the problems because the relevant behavior (rise in the price of masks) had not been scientifically explained. As a result, government would impose such a counterproductive policy as price control.

The theory of labor value only regards labor as the only source of the value of a good, which has been clearly refuted by the fact that a landlord who does not work at all can collect rent. Marx defined the collection of rent by a landlord as "exploitation", which meant he still insisted that all income from land should only be distributed to labor, and just a part of the income was exploited by the landlord in the form of rent. Marx lived in the era of industrialization, so when he further extended this "exploitation theory" to industrial production, he naturally deduced the "theory of surplus value" that capitalist exploit workers.

The concept of exploitation is full of subjective value judgment (moral judgment). And if it is regarded as truth, the policy proposal must be prohibition of landlords from exploiting farmers, which means landlords are not allowed to collect rent, so that farmers get all the income from lands. However, if landlords could not get any income, they must take back their lands and not lease to farmers as a countermeasure. Then how could agricultural production be carried out without lands and only with the labor of farmers? As a result, if one insisted on the theory of labor value in classical economics, he must also insist on the exploitation theory and the theory of surplus value of Marx, and finally he would logically come to the conclusion that he must advocate "public ownership" which completely deprive landlords of their PPRs of lands and transfer to the nation. Then lands are not owned by specific individuals, but owned by an ambiguous subject, which ensures that lands will be input to agricultural production even if no rent could be collected.

However, public ownership completely violates PPRs, and completely eliminates the market or the price as competition criterion, thus leading to huge rent dissipation (one kind of transaction costs), which results in economy recession and even collapse. Self-interested people have to give up public ownership and return to private ownership in order to reduce transaction cost and survive. One can refuse to face the reality like ostriches, but the reality will not change because he pretends not to see it. Economic laws are the same as physical laws, regardless of whether one likes or not. Whoever does not abide by them will be severely punished.

Back to the concept of rent. The reason why a landlord can collect rent is because the land he owns does contribute to production and is one of the sources of value for the output (agricultural products). Therefore, part of the income from the land must be distributed to the land, actually to the owner of the land. Thus, rent is the income distributed to the land for its contribution to the agricultural production.

Later, Marshall discovered that not only the income of land is rent, but the incomes of all kinds of production factors are also rents. For example, the income of the so-called capitalist does not come from exploitation of the surplus value of workers as Marx said, but from the provisions of two types of production factors.

Firstly, he provides capital, which means he lends his money to himself for production, so he requires relevant income. Seen from the concept of opportunity cost, the essence of it is rather easy to understand. If he does not use the money for production but deposits it in a bank, he has interest as income. Now he uses the money for production, thus giving up the interest of saving, which is his opportunity cost. In fact, when he deposits the money in a bank, the bank just play the role of intermediary to help him lend the money to others who need funds. Therefore, the interest he gives up is the income he can obtain by lending the money to others for production. Seen from another perspective, if he does not use his own money, he can borrow money from a bank, and needs to pay interest. In a word, using one's own money is actually borrowing money. Just because he borrows money from himself, he pays the interest to himself.

Secondly, the so-called capitalist provides his own labor in management and other aspects, which is mainly mental labor. Blue-collar workers provide mainly manual labor, while white-collar workers provide mainly mental labor. In essence, he rents his labor to himself for production, so he requires relevant income. Seen from the concept of opportunity cost, the essence of it is also rather easy to understand. What he gives up is the wage of renting his own labor to others for production. Therefore, the difference between working for others and self-employment does not lie in whether the labor is rented out or not, but only in whom the salary is paid to.

In Marx's era, the owner and the manager of a firm were the same person, so Marx could not see clearly that this person actually had played two roles at the same time, and had two incomes at the same time: one is the income of interest as the owner of his own capital, the other is the income of wage as the owner of his own labor. It is somewhat logically similar the situation

mentioned in the previous lecture that when there is information cost, historical cost and opportunity cost cannot be clearly separated, so economists misunderstood that people made decisions according to historical cost. Nowadays, corporation has become a very popular form of firm. It is clear that shareholders (the owners of the firm, or strictly speaking, the owners of capital) and managers (strictly speaking, the owners of mental labor) can be two different groups of persons. Shareholders receive dividends, while managers receive salaries, so the owners of different production factors all have their own income. Thus, the income of various production factors is actually the rent paid for leasing them for production.

Since the income of all production factors is their own rent, Marshall generalized the term of "rent" from a narrow sense which only referred to the rent of land to a broad sense which refers to the income of all production factors. At first, in order to distinguish from the narrow sense of rent that people were used to understand, he called the rent in broad sense as "economic rent" or "quasi-rent". As many developed countries have developed from agricultural economy to industrial economy, the concept of rent is generally recognized in economics as referring to the broad sense.

Look back on the example of the options of a piece of land. There are only two options "Use" and "Not use" when the most roughly observed. It is obvious that "rent of land is the income of land" is actually a particular case of "rent is the income of production factors".

As will be explained in the later lecture on "Interest Theory", any that can bring in income is an asset. The purpose of inputting factors to production is to share a part of the output, so production factors are all assets. And assets are valuable because they can contribute to production and get income from it. As mentioned in previous section, income is rent, so rent is the value of the asset. In fact, all that human need is asset, no matter whether they are needed for production or consumption. Consumer goods can also be regarded as production factors, because they can be regarded as reproduction of the labor capacity in human. Thus, all that human need is valuable, can bring in income and has rent. Fundamentally speaking, the value of it is determined by the two constraints of self-interest and scarcity which jointly determine the price of it, and the price constitutes the income for the producers, so the relevant income is the rent of it.

We should be flexible in thinking and observe the same phenomenon from different angles, and then we can really understand how the concepts in economics are connected with each other.

Section 4: The Concept of Rent from the Second Perspective

Seen from the second perspective, rent is the part of income when the supply does not change with the income. This definition sounds very difficult to understand, let's give an example. Classical economists noticed an obvious feature of land that the supply of it seemed unchanged. There is so much land on earth, which will not change due to the change in rent of land (income from land). Therefore, rent seems to be a "surplus" of income, which means even if the rent of land falls to zero, there is still so much land available on earth. Some of classical economists such as Henry George advocated that government should impose all taxes on landlords, so that even if their income of rent dropped to 0, the supply of land would not be affected, thus neither would the whole economy be affected.

However, is it really the case? As mentioned in the previous section, if a landlord is not allowed to obtain rent of land based on the exploitation theory of Marx, he will take back the land from the agricultural production. In other words, if the rent of land is really reduced to 0, landlords will choose not to use the land, so that although there is much land on earth, the amount of land that is effectively supplied to the market is 0.

What if the rent significantly decreases but does not drop to 0? It is similar to the effect of price control which is strongly enforced, and effectively prohibits of all kinds of means such as contract fee from going around it. Suppose the price control stipulates the rent of a house cannot be more than RMB800 a month, the houses originally worth more than RMB800 would be withdrawn from the house renting market, and switched to the factory or shop leasing market or simply sold out. In other words, a piece of land or a house can be supplied not only to the market of a specific production, so the supply of it will decrease with the decrease of rent in that specific market.

What is more, a landlord can also reduce the quality of his house to a level compatible with the regulated price of RMB800 such as no provision of furniture and no maintenance of the house. The houses originally worth RMB1000 are still supplied to the house renting market under the rent control of RMB800, which seems that RMB200 of rent is a surplus. However, does the supply of landlords really not change? No! The supply of the houses with high quality which are so high-priced decreases completely to 0! Only the houses with low quality compatible with the regulated rent of RMB800 are supplied to the market.

Here is another example of rent decrease not caused by price control, because it is more universal and there is not additional effect of rent dissipation caused by price control. As we all know, there are many houses around a university for students to rent, and there is obvious seasonal fluctuation of rent. The beginning of each semester is the peak season for the houses, so the rent will rise to the highest at that time. And the end of each semester, especially the Spring Festival in winter vacation, is the off season for the houses, so the rent will drop to the lowest at this time. Suppose the rent in peak season is RMB1000, while that in off season is RMB800. The rent has changed, but the supply of the houses does not change. The difference of RMB200 seems to be a surplus. However, although the supply of the houses remains unchanged, do the landlords really not change at all? Of course not. The landlords in peak season will be more active in cleaning their houses, and may even spend money on renovating the houses, so as to increase the quality of the houses to the level that worth the rent of RMB1000. By contrast, in off season, landlords will be less active in maintaining their houses in the best condition to attract tenants. In other words, the quantity supplied remains unchanged, while the quality supplied has changed.

The thought that change in the income of land will not cause change in the supply of it, just because the options are observed roughly, so that only two options of "Use" and "Not use" are observed. It seems as if the supply of land will not change until the rent falls to 0. However, if the option of "Use" is further subdivided, there are in fact many sub-options included in it, such as used as "agricultural land", "industrial land" and "residential land". For example, if the income from agricultural land falls, some of the lands will be withdrawn from the market of agricultural land and turn to that of other uses of land. Of course, there are still some lands unchanged in the supply of agricultural land. However, there are also in fact many sub-options in the option of agricultural land, such as growing different crops, with or without fertilizer, and so on. The agricultural lands have actually changed to some extent along with the change in the income, but just have not changed beyond the scope of the "agricultural land". Thus they seem to have not changed, and there seems to be part of income when the supply does not change with the income, which is precisely rent. In other words, it is again the amazing feature of rent that it seems sometimes large and sometimes small and sometimes zero depending on how the options are observed.

There is rent not because the behaviors such as supply have not changed, but the changing options are not observed. Remind the statement in section 2: All income is cost, only part of it is the cost of the options observed, and part of it is the cost of the options not observed. In order to distinguish the options observed and not observed, the former is directly called "cost" and the latter is called "rent". Now it should be easier to understand this second perspective from which the definition of rent is the part of income when the supply does not change with the income. The part of income where the supply is observed to change with the income has classified as cost, while the part of income where the supply is not observed to change with the income is regarded as unchanged and classified as rent.

The rent of land can be generalized to that of all other production factors or assets. Take Elvis Presley as the example of labor. Elvis Presley was a truck driver before he became a popular singer, and earned just a few hundred dollars a month at that time. By contrast, he earned tens of millions of dollars every year as a singer. For the sake of simplicity, suppose he earned

1,000 a year when he was a driver, while he earned 10 million a year when he was a singer. As shown in Table 12-6, the rent of Elvis Presley as a singer was as high as 9.999 million, which means he would not change from a singer to a driver until his income as a singer dropped from 10 million to \$1,000, so he seemed to be unchanged in supply of singer.

	Income	Cost	Rent
A: Driver	0.001	10	-9.999
B: Singer	10	0.001	9.999

(Unit: Million US Dollar)

Table 12-6

However, it is only because the options of Elvis Presley are observed too roughly, and he is thought to have only two options: "Driver" and "Singer". In fact, the option of "Singer" can be subdivided into numerous sub-options, such as choosing different record companies, choosing different songs, choosing to release albums or to hold concerts, choosing to sing more or less songs, even choosing to laugh more or less ... all are different sub-options. If all the sub-options are sorted by their incomes, the income of the sub-optimum must be infinitely close to that of the optimum, thus is the cost of the optimum, and the rent of the optimum is infinitely close to 0.

In fact, the case shown in Table 12-6 is not the roughest observation of Elvis Presley's options. Both "Driver" and "Singer" are jobs, and are the same option if observed most roughly. There are only options of "Work" and "Not work" if observed most roughly, which is shown in Table 12-7 that Elvis Presley's rent reached the maximum of \$10 million when all his income was rent. Is it essentially the

same as the options of "Use" and "Not use" for a piece of land? Thus, not only the supply of land does not change with income, but it is also the same for labor, actually for all production factors or assets. Classical economists thought that the feature of the rent of land is unique, but is actually universal.

	Income	Cost	Rent
A: Driver	0	10	-10
B: Singer	10	0	10

(Unit: Million US Dollar)

Table 12-7

After explaining the two perspectives for understanding the concept of rent, readers should find them the same in essence.

From the first perspective, rent is the income of production factor or the value of asset, which is the roughest observation of options, so all the income of production factor or asset is rent. From the second perspective, options are not sub-divided into countless sub-options, so some options have not been observed, and it seems that the supply does not change with the income, and part of income is regarded as rent.

Although the two perspectives are essentially the same, it is still necessary to distinguish from which perspective the concept of rent is used in a specific case. For example, the "rent" in "rent dissipation" refers to the concept from the first perspective, because it means the loss in the income of an asset when the price of it is regulated. And the concept of overhead cost to be introduced in the next lecture is essentially rent from the second perspective, because it means that income fluctuation in an industry does not necessarily cause producers to leave the market and reduce supply.

Section 5: Profit

As mentioned at the beginning of this lecture, it is inappropriate for MSE to use the term "profit" to refer to the difference between income and (opportunity) cost. After the explanation of the concept of rent in previous sections, readers should understand the concept of rent is far beyond the scope of that of profit. In essence, profit in MSE is a concept defined by a mathematical formula, which is not as rich in economic implication as rent.

More importantly, the term "profit" needs to be defined differently from rent in this book and profit in MSE. We define "profit" as the difference between actual income and opportunity cost.

As has stressed in the previous lecture of "the Concept of Cost", cost in economics must refer to that in the future, which is completely different from the actual expenditure (historical cost or accounting cost) in the past. In other words, cost must be expected cost. Expectation is intention, not a fact, but it can be deduced by examining the objective facts that make the expectation come into being. It is logically similar that income in economics is expected income, not actual one. Before making a choice, one expects the future, which means he expects the income and cost in the future, and the difference between them is rent, and then he makes a choice according to the principle of rent maximization. When the choice has been made, the difference between actual income and expected cost (opportunity cost) is profit.

Based on the previous lecture and this lecture, there are three formulas: (1) rent = expected income - expected cost (opportunity cost); (2) accounting profit = actual income - actual cost (historical cost/accounting cost); (3) profit = actual income - expected cost (opportunity cost). As mentioned before, all income is cost, and there is not such income as is not cost, so expected income = expected cost (opportunity cost) as long as options are sub-divided into countless sub-options. Thus, the formula of profit can be rewritten as: profit = actual income - expected income. In other words, profit is essentially unexpected income.

Once a choice is made, it is like the water poured out, which means the actual income, just like historical cost, will no longer affect people's choice. Similarly, profit cannot affect people's choice, because whether there is profit and how much it is cannot be expected. People choose according to expectation, so profit beyond expectation is meaningless.

For example, if one accidentally picks up a wallet when he walks on a road, it is his unexpected income or profit. If one accidentally loses his wallet instead, it is also his unexpected (negative) income or (negative) profit. Thus profit can be positive or negative.

There is profit because there is information cost which makes one's expected income different from actual income: he does not know in advance that he can pick or lose wallet on the road. As long as the constraint of information cost remains unchanged, profit cannot affect people's behavior. One will not walk on that road every day after accidentally picking a wallet and expects to pick a wallet again. Neither will he avoid walking on that road just because he has accidentally lost his wallet and expects to lose it again.

Lecture 13: Direct Cost and Overhead Cost

Section 1: The Fallacy of Fixed Cost

In MSE, it is said that cost is the sum of fixed cost and variable cost. The so-called "fixed cost" refers to the cost of fixed input that cannot be changed in short run, while "variable cost" refers to the cost of variable input. However, what is "short run"? It refers to the period when part of cost is fixed, which means there is fixed cost, while "long run" refers to the period when all costs are variable. There is obvious logical problem of circular reasoning in MSE about fixed cost.

Economists cannot clearly say how short the short-run is, but that it is short-run if producers have not changed all the inputs, while fixed cost is defined based on the period is too short for the inputs to be changed, which is in fact a tautology.

What is more, there is also a fatal fallacy for the concept of fixed cost. Producers cannot change the fixed inputs in short run, which means they do not have the option of changing the fixed inputs in short run. However, as has stressed in the lecture of "the Concept of Cost", there is no cost if there is no alternative option. Since there is no option of changing the fixed inputs, how can there be cost?

Essentially fixed cost is historical cost, which comes from the production factors called "fixed assets" in accounting that need to be purchased or rented before production. However, historical cost is not cost. Before these production factors are purchased or rented, there are costs for them. Once they have been purchased or rented and the relevant price or rent is paid, these expenditures become historical costs, and are not costs any longer.

It is the most serious fallacy in the theory of firm in MSE: on the one hand, at the very beginning, it is stressed that cost in economics is opportunity cost, not historical cost; on the other hand, when it comes to cost function (and hence production function) later on, fixed cost which is essentially historical cost is put into the cost function, and it seems that the concept of opportunity cost is completely forgotten.¹

If we insist that cost is opportunity cost, not historical cost, what are the costs of the so-called fixed assets that have been purchased or rented before production? Obviously, the costs of these production factors after they have been purchased or rented should be the income given up if they are resold or sublet, which is an alternative to retaining these production factors.

If we insist that cost is opportunity cost, not historical cost, what are the costs of the so-called fixed assets that have been purchased or rented before production? Obviously, the costs of these production factors after they have been purchased or rented should be the income given up if they are resold or sublet, which is an alternative to retaining these production factors.

For example, when a machine breaks down and stops production for one hour, except that electricity can be saved, it seems that all other inputs cannot be changed, and belong to fixed input, the cost of which fixed cost. However, is it really impossible to change these inputs? In principle, the boss can immediately dismiss all workers, and resell or sublet other machines, equipments or even the buildings. Of course, it is crazy to do so for just one hour of production suspension. Yes, producers will not do so, but it is just that they do not choose to do so, not that they have no such options! There is no cost if there is no option, while there is cost if there is such an option which is not chosen.

Why is the option not chosen? Because it is too costly. It is very troublesome to dismiss all workers, resell or sublet other machines, equipments and buildings within one hour. And it is even more troublesome to buy or rent back all the necessary production factors again when production resumes in an hour. The description "to be troublesome" in daily use means "there is

high transaction cost". The benefit of paying such a high transaction cost is only to save the waste caused by these production factors which have been idling for one hour, which is not worth it. However, if the production suspension lasts for one day, the boss will choose to dismiss temporary workers, because they are paid on a daily basis, and the transaction cost for dismissal and re-recruitment is very low. It is less costly compared with the alternative of paying their wages even they need not work.

When the production suspension lasts for a month, the transaction cost for dismissal and re-recruitment of workers employed on a monthly basis also becomes lower than the cost of keeping them. It is the same for the machines and equipments leased on a monthly basis.

Finally, if the boss plans to permanently shut down, there is no need to buy or rent back all the production factors again, and the transaction cost for resale or sublet machines, equipments and buildings also drops to a level worth being chosen.

The above example only shows the case of production reduction, and it is similar for the case of production increase. For example, a boss of restaurant will only employ part-time workers to help during the peak hours of meal time, and will not increase the area of the restaurant. It seems that labor is variable factor, while the area of the restaurant is fixed factor. But the reason is just that the transaction cost of changing the area of the restaurant is obviously higher than that of changing the employment of part-time workers.

If a restaurant continues to thrive, the boss is sure to expand the area. There has been a small porridge shop near my home. Because the porridge is very delicious and the price is rather cheap, it has become so popular with the residents nearby, that there are much more customers than the tables in the peak time of the morning, and many of them have to stand outside the shop to have porridge. At first the boss only rented one shop front, and then he rents another shop front beside it to increase the business area.

Thus, it is not that producers have no options, but there is transaction cost, which makes it more difficult for some production factors to change immediately along with the changes in output than others. The transaction cost is precisely the cost of choosing to change. If it is too high, producers will not choose to change.

It is Marshall who invented the concepts of short-run and long-run term in MSE. He actually felt that there was transaction cost of changing inputs, but he did not even know the correct concept of opportunity cost. There is no concept of transaction cost in economics until Coase discovered it almost 50 years later. So Marshall invented these two concepts to avoid using the concept of transaction cost to explain producers' behaviors of choosing to change some inputs but choosing not to change other inputs. However, as mentioned earlier, there is logical problem of circular reasoning for these two concepts and fixed cost, and it is essentially a tautology that never really explains producers' behaviors. What is more, the concept of fixed cost is essentially historical cost, which is out of line with the concept of opportunity cost.

1.

Why is there such obvious inconsistency in MSE? The reason is that there had been the theory of the cost of production before the concept of opportunity cost. It is Marshall who came up with the theory about the cost of production, but there had not been the correct concept of opportunity cost in economics at that time, so the wrong concept of historical cost was still prevalent. Therefore, when Marshall thought about the cost of production, he still used the concept of historical cost, not opportunity cost at all.

As pointed out in the previous two lectures, there are two mistakes in the theory of labor value: one is that labor is regarded as the only source of the value of a good, and the other is that historical cost (the labor input during the production) is cost. Marshall had realized the former mistake in the theory of labor value, and found that other factors besides labor also contribute to production and create value, so they should also have relevant income in return. However, he did not fully realize the latter mistake in the theory of labor value, and continued to bring this mistake into the theory about the cost of production.

When opportunity cost has been accepted as the correct concept of cost in economics, economists should revise all the theories came up with before the concept of opportunity cost including the theory about the cost of production with the wrong concept of fixed cost. However, economists have not done so, which causes that there is such obvious inconsistency between the concept of opportunity cost and fixed cost in MSE. Now we are doing so to eliminate any concept or theory inconsistent with the concept of opportunity cost.

Section 2: Direct Cost and Overhead Cost

Abandoning the concepts of short-run and long-run, and facing the reality where there is transaction cost, how should we analyze the cost of production? A more objective criterion can be used to divide it into two kinds: one is the cost when there is output and it changes along with the change of output, which is called "direct cost" (DC), and is equivalent to variable cost in MSE. The other is the cost even when there is no output, which is called "overhead cost" (OC), but it is different from fixed cost in MSE.

There are three statements about overhead cost. (1) The definition of overhead cost is: when producers choose to continue to do business, the cost that they have to bear even without production is called overhead cost. (2) The calculation formula of overhead cost is: income - direct cost. (3) The essence of overhead cost is: the asset value of the whole business, i.e. the rent. These three statements can be deduced from each other. The following is an example of a training agency.

Suppose one is considering whether to do the business of a training agency. For the sake of simplicity, suppose there are only two production factors needed: rooms for classrooms and electricity. Labor (including the boss, recruiters, and teachers, etc.) is excluded here because it is a very complex production factor, the cost of which is not easy to distinguish as direct cost or overhead cost. Therefore, although it is impossible not to use labor in any production, in order to prevent beginners from confusing, let us simplify the assumption to so unrealistic first. Later, we will extend the assumption to more realistic.

Firstly, one investigates the market and finds that once he enters the market of training agency, only when there are courses, there will be the cost of electricity, because he turns on the lights and computers during a course, and can turn off them if there is no course to avoid such costs. Thus, electricity charge is the direct cost after entering the market. Suppose according to the investigation, if there are courses every night, per month the income is RMB1000, the electricity charge will be RMB200, and the rent for the classrooms is RMB800. In other words, income can cover all costs, so he decides to enter the market.

Someone may ask: income only covers all costs, which means there is no profit, why will one choose to enter the market? If one does not enter a market unless there is profit, in fact there are other production factors such as labor, capital, etc. The so-called profit is used to cover the cost of these other production factors. As mentioned in Lecture12, one who works for himself and use his own money also needs to pay wage and interest for himself, so the profit is actually not profit, but just the wage and interest paid to cover the cost of his own labor and money. And in the example of training agency here, for the sake of simplicity, it is supposed only two production factors (room and electricity) are needed. In other words, other production factors including the labor and money of the boss himself are assumed not to be needed or to be free, so he does not need so-called profit to cover all these costs. It is called zero-profit theorem in economics which will be introduced in the later lecture.

Before entering the market, all costs (rent of classrooms and electricity charge) are direct costs, because one can choose not to enter, and does not need to bear these costs, which means there will be no such cost if there is no production.

And then, one has paid the rent of classrooms for one year, and the cost equivalent to RMB800 a month will immediately become historical cost. In principle, after entering the market, as long as the income from the market is higher than the direct cost of electricity (RMB200), one will continue to do business, because the rent of RMB800 has become historical cost which should not affect his behavior. However, if one expects the income from the market only more than RMB200 but less than RMB1000 before entering the market, he will not choose to enter, because at that time, the rent of RMB800 is still direct cost, not historical cost.

After entering the market, one may find the income of tuition is RMB1000 as expected. After deducting the electricity charge of RMB200, the net income is RMB800 which can cover the rent paid when entering the market. This net income is precisely overhead cost according to the calculation formula of "overhead cost = income - direct cost". It seems that overhead cost is the same as fixed cost which is historical cost paid when entering the market, but it is only equal in value and completely different in essence. The essence of fixed cost is historical cost, not (opportunity) cost, while the essence of overhead cost is rent which is (opportunity) cost.

Why? Suppose that one night there is no course for some reason, one will turn off the lights and the computers to save the direct cost of electricity. But how about the cost of the classrooms? Regardless of the relevant transaction cost, theoretically one can choose to sublet the unused classrooms on that night, and the rent he gets is the opportunity cost of the classrooms. Thus, the rent of subletting the classrooms is the cost even when there is no production (the rooms are unused), which is the overhead cost.

The question is: how much will someone else be willing to offer to rent the rooms? The answer is: it must be equivalent to RMB800/30 days, because it is only one night's rent and RMB800 is the rent of a month. The rent someone else is willing to pay cannot be higher than the income from using the rooms. Suppose someone else also uses the rooms as classrooms for training agency, exactly the same as the lessor. The rooms have been arranged to be suitable as classrooms for training agency, so it is better to use them in the same way, otherwise there will be additional cost. In other words, the option consistent with income maximization or cost minimization is to use the rooms in the same way as the lessor, which is to use them as the classrooms for training agency. Since the rooms are used in the same way, the income must be also the same, which is equivalent to RMB800/30 days or the difference of income (RMB1000) and direct cost of electricity charge (RMB200).

Thus, the concept of overhead cost is consistent with the calculation formula. When there is no output, the income from transferring an idle asset is the opportunity cost of not transferring it. And the price at which someone else is willing to buy it is the indifference between the income the lessee can get and the direct cost. Since the best way for the lessee to use the asset in the same way as the lessor, his income is also the same as that of the lessor. Therefore, the overhead cost of the asset is the difference between the income and the direct cost.

Caution: income, direct cost and overhead cost are all expected value, not actual one. If it has happened, it has become history. Take the business of training agency as example, when one has got RMB1000 as the income, and has paid RMB200 as the electricity charge, so has earned 800RMB as the net income which can make up for the rent of the rooms in the first month, all of these are no longer (expected) income, (expected) direct cost and (expected) overhead cost. Now one needs to estimate the income and electricity charge of the next month. The business in the first month has strengthened his expectation, so he will continue to expect the income of the next month to be also RMB1000, the electricity charge RMB200, and the net income RMB800, all of which are (expected) income, (expected) direct cost and (expected) overhead cost.

As mentioned earlier, overhead cost appears to be equal in value to fixed cost (historical cost) just when the expectation is correct. If the expectation is incorrect, overhead cost will be no longer equal to fixed cost. For example, suppose the income in the first month is RMB900 instead of RMB1000, and the electricity charge is still RMB200, so the net income becomes RMB700. The actual income (RMB900) is lower than the expectation (RMB1000), so there is RMB100 of (negative) profit which is unexpected (negative) income.

Now one still needs to estimate the income and electricity charge of the next month. If he thinks that the loss of RMB100 in this month is just an accidental loss which will not last in the next month, he will continue to expect the income of the next month to be also RMB1000, the electricity charge RMB200, and the net income RMB800, which means the overhead cost is still RMB800, not RMB700! Here it is very clear that overhead cost is not the difference between actual income and direct cost, but (expected) income and direct cost.

However, if one thinks the actual income lower than expected one not because of accidental bad luck but inaccurate information about the market, he should adjust his estimate about the situation in the next month. Suppose one expects the income of the next month to be RMB900, the electricity charge is still RMB200, and the net income or overhead cost RMB700.

Readers can deduce the contrary situation when the performance is better than expected by imitating the above analysis.

Thus it is clear that if one adjusts his expectation, overhead cost will be different from fixed cost, and can change along with the change in expectation. By contrast, fixed cost is, of course, fixed, and will never change, because it is essentially historical cost, and what has become history will never change again.

The analysis above is the most simplified, so next we will extend the analysis by relaxing the assumptions.

Suppose there are more than two production factors, they can be classified into those similar to rooms and those similar to electricity. The costs of the production factors similar to electricity are all direct costs, and the costs of the production factors similar to rooms are all overhead costs. When there is no output, instead of transferring only an idle production factor (such as room), the whole business (excluding production factors similar to electricity) is transferred, and the income will be the rent of the whole business. Thus, this rent is the overhead cost of the whole business, which is shown in Table 13-1.

	Income	Cost	Rent
A: Shutdown	0	Income - DC	- OC
B: No Shutdown	Income - DC	0	OC

Table 13-1

As has been explained in the previous lecture, there will be rent if the options are observed roughly. Overhead cost is rent, because only two options are roughly observed: shutdown (to withdraw from the market) and no shutdown (to stay in the market and continue the business). Obviously, as long as there is overhead cost, even if the income is less than expectation, but at least can still cover the direct cost, producer will not choose to shut down. In other words, overhead cost can be regarded as rent from the second perspective which means the part of income when the supply does not change with the income. Here even if the income from the market has decreased to be less than expectation, producer still chooses not to shut down or stay in the market to continue the business, so the supply does not change. It is why the essence of overhead cost is rent.

To sum up, the definition, the calculation formula and the essence of overhead cost can be deduced from each other.

Section 3: The Explanation of Producers' Behaviors with Overhead Cost

The concept of overhead cost is very important because it can explain many producers' behaviors.

Firstly, it is overhead cost that determines the shutdown decision of producers.

As can be seen from the analysis in the previous section, even if income (RMB900) is lower than expectation (RMB1000), producers will not choose to shutdown, unless income falls so low that it cannot cover direct cost (RMB200). In other words, whether producers stay in the market and continue the business depends on overhead cost.

In reality, income always fluctuates, but producers will not shutdown the business frequently. As long as one expects income will not decline to be lower than direct cost, he will stay in the market, and what he loses is part overhead cost, which is the "pain" he can bear. However, once income drops to lower than direct cost, producers will shut down immediately.¹ For an analogy, direct cost is like bone, while overhead cost is like fat. Producers experience economic recession just as animals experience winter. The thicker the fat (the higher the overhead cost), the longer animals can survive (the longer producers can survive). And once the knife cuts into the bone (income falls to be lower than direct cost), animals will die (producers will shut down business or be knocked out of the market).

Secondly, logically deduced from the first key point, in order not to be knocked out of the market, producers must try to maximize overhead cost. The zero-profit theorem mentioned in the previous section determines that producers actually cannot have profit. In MSE, it is said that what producers try to maximize is profit, which is completely wrong and again is a big mistake of MSE.

It sounds ridiculous that the target of producers is overhead cost maximization, because cost seems to ought to be minimized. However, the essence of overhead cost is rent, and rent maximization is rather reasonable. To be precise, the "cost" in cost minimization should refer to direct cost, not overhead cost. As mentioned in Lecture 12, all income is cost, and there is not such income as is not cost. If cost minimization means that all costs are minimized, is it not similar to income minimization? Obviously, it is not correct. In order to survive the fierce and cruel market competition, producers must strive for as much overhead cost as possible to protect them. Ceteris Paribus, producers with higher overhead cost can survive longer in economic or industrial recession with income declining.

To increase overhead cost, there are two directions: one is to reduce direct cost as much as possible, and the other is to increase income as much as possible, which is subject to competition constraints from other rivals in the market and out of the market who intend to enter the market at any time.

There are many behaviors of producers to increase overhead cost. For example, why does a firm spend a lot of money on R&D (Research and Development) and innovation which increase cost? Because there are two functions for technology: one is to reduce production cost which is direct cost, so that producers can get higher profit (actually overhead cost) with the same price (income). The other is to improve the quality of products, so that consumers are willing to buy even with higher price. Thus the money invested in R&D is used to increase overhead cost which here is the rent of technology.

Why does a firm pay high salary to employ an able executive? Because an able executive is either a cost-killer who can drastically cut cost (direct cost, or a marketing expert who can greatly increase the sales of products in the market. Thus the money paid to hire an able executive is also used to increase overhead cost which here is the rent of human capital of the able executive.

Why does a firm spend a lot of money on advertisement to build brand? Because the essence of brand is to reduce the information cost of consumers on the quality of the products, so that consumers are willing to buy brand-name products even with higher price. Thus the money invested in building brand is also used to increase overhead cost which here is the rent of brand.

Thirdly, overhead cost can protect producers from being knocked out of the market by competition, while market competition can in turn protect overhead cost. The income that producers can get from selling the whole business is the direct cost of those who are still outside the market and intend to enter. As long as there are outsiders who intend to enter the market at any time, one of their options is to directly purchase the whole business transferred by insiders. Therefore, these outsiders who compete to enter the market have objectively protected the overhead cost of insiders from falling to zero. Competition will dissipate profit, but will protect rent which is overhead cost. Precisely speaking, the competition among the outsiders to enter the market can protect the overhead cost of insiders which will in turn protect the insiders from being knocked out of the market. The direct cost of insiders is naturally lower than that of outsiders, so the insiders have the advantage of cost in the competition with outsiders.

How does market competition protect the overhead cost? There are two aspects. Suppose one has store, and per month the income is RMB12,000, the direct cost of water, electricity, labor, etc. is RMB2,000, so the overhead cost is RMB10,000 which is the rent of the whole business.

Theoretically, once entering the market, one will not choose to shutdown unless the income falls to be lower than RMB2,000, as shown in Table 13-2.

	Income	Cost	Rent
A: Shutdown	0	10000	-10000
B: No Shutdown	12,000 - 2,000	0	10000

Table 13-2

However, why does the income fall? The analysis in detail indicates two different possibilities.

One possibility is that the insider is poor in doing business, and the income is less than RMB 12,000 which is common for other insiders. Suppose his income is only RMB10,000, which causes his overhead cost to fall to RMB8,000. Theoretically, he will continue to do business in the market, because there is still overhead cost. As long as the income does not drop to be lower than RMB2,000 or the overhead cost does not drop to 0, it is the "pain" he can bear.

Really? No! Outsiders will find it profitable to offer this insider a price of not less than RMB8,000 but not more than RMB10,000 to buy the whole business (the store). Suppose an outsider offer RMB9,000. Will the insider accept it? Yes! Because now his options become that shown in table 13-3:

	Income	Cost	Rent
A: Shutdown	9000	8000	1000
B: No Shutdown	10,000 - 2,000	9000	-1000

Table 13-3

Attention: now the income of "Shutdown" is no longer 0, because the insider can get RMB9,000 by selling the store to the outsider when he chooses to shutdown, which leads to an increase in the opportunity cost of "No Shutdown", and he becomes inclined to choose to shut down. If the income of "Shutdown" is adjusted to 0 so as to be comparable with the data in table 13-2, it is shown in table 13-4. It is clear that the competition from outsiders has changed part of the overhead cost of the insider to direct cost, so the overhead cost have decreased, which makes the insider tend to choose to shut down.

	Income	Cost	Rent
A: Shutdown	0	-1000	1000
B: No Shutdown	10,000 - 2,000 - 9,000	0	-1000

Table 13-4

What is more, the overhead cost of the insider has decreased, but after the outsider have entered the market by buying the store from the insider, the overhead cost of the store will rise back to RMB10,000 ($= 12,000 - 2000$). In other words, market competition has protected the overhead cost of the store from falling to RMB8,000 or even to the theoretical minimum of 0 caused by the poor performance of insider.

This numerical example clearly shows how the market competition protects overhead cost of the store from being lower than the normal level of the whole industry, and allocates resources to the most efficient producers by the survival of the fittest. For the insider, he has made the optimal choice by selling the store to the outsider, because he can get RMB9,000 which is more than RMB8,000 if he chooses to continue the business. For the whole industry, this insider with poor performance has been knocked out of the market by the outsider, and the resource of the store has been transferred from the less efficient insider to the more efficient outsider through price competition. It ensures the most effective use of the resource, and protects the overhead cost (i.e. the rent or the value) of the resource.

It is even clearer if we observe public companies in the stock market. If the performance of the CEO in a public company is so poor that the stock price drop dramatically, outsiders will buy the stocks in large quantities. When they can control the company, they will enter the board of directors and fire the incompetent CEO, thus protecting the equity (overhead cost) from falling to 0 or even going bankrupt.

The other possibility of income decline is that the business environment of the whole industry is worsening, which is not the problem of a certain producer's poor performance. Suppose the income of all producers generally drops from RMB12,000 to RMB10,000, causing the overhead cost of insiders to drop from RMB10,000 to RMB8,000. Now there will be no outsiders who are attracted to enter the market, because they also expect to get only RMB10,000 instead of RMB12,000. For outsiders, all costs are direct costs which are RMB10,000, while the direct costs of insiders are only RMB2,000 which are much lower, so outsiders cannot compete with insiders. In other words, with the same other factors (such as the ability of doing business), the direct cost of insider is lower than outsider because outsider does not have overhead cost and all costs are direct costs. Thus insiders are protected by overhead cost and free from outsiders' competition before the income falls to be lower than overhead cost.

In brief, the competition of outsider' entering the market protects the overhead cost of the assets held by insiders, which in turn protects insiders from the competition of outsiders. Of course, the premise of the protection is that insiders are not less competitive than outsiders. In other words, the competitiveness of an insider is precisely reflected in the overhead cost of the assets he

holds. The more competitive he is, the higher the overhead cost of his assets, and the harder it is for outsiders to enter the market to compete with him and knock him out.

What is more, the income decline of the whole industry may be due to other reasons except for the economic recession. For example, there may be technological progress or other factors that leads to the decline in production cost or transaction cost. There are also two possibilities.

One possibility is that new industry is emerging, which is a perfect substitute for the old industry, and more competitive than the old one. The old industry may be replaced or even completely eliminated by the new industry. The most typical example is digital camera replacing film camera. In fact, it is logically similar to an insider with poor performance is replaced by a more competitive outsider. Therefore, when the income of the old industry decline, market competition will not protect the overhead cost of it, but will protect the overhead cost of the whole economy! Market competition will replace the overhead cost of the old industry with that of the new industry to ensure that the overhead cost of the whole economy is protected. It is similar to the previous analysis of market competition which replaces the poor-performed insider (equivalent to the old industry) with the competent outsider (equivalent to the new industry) to protect the overhead cost of the store (equivalent to the whole economy).

The other possibility is that some factors (such as technological progress) lead to a drop in production cost or transaction cost which is the entry threshold of an industry, resulting in a drop in the cost of entry for outsiders. The entry threshold is part of the overhead cost of insiders, which is monopolistic rent, so the lower threshold will cause the overhead cost of insiders to fall.

One of the typical examples is the development of microfilm. In the past, movie-making required high technology and much capital, which means that the entry threshold of the movie industry was very high. However, the progress of Internet technology has given birth to microfilm, of which the requirement of technology and capital for making movie is much lower than before. Many directors are indignant, because they used to spend a lot of energy and money on movie-making, while young people now can easily make microfilm. They accuse the microfilm of being too low in threshold and chaotic. They had spent a lot of money when they entered the movie industry. The cost they had paid is historical cost, and has converted to overhead cost which is covered by the high price of their products (movies) due to the high entry threshold keeping outsiders away. Now young people can easily make microfilm, which means the entry threshold for outsiders has been greatly reduced, resulting in a significant drop in the overhead cost of senior directors.

It is profit dissipation caused by fiercer market competition, which seems somewhat similar to rent dissipation. Rent dissipation is due to infringement of PPRs, while profit dissipation is due to the factor change (such as technological progress) causing the monopolistic rent turn to profit which will be surely dissipated by market competition. When the entry threshold or monopoly of an industry declines and competition becomes fiercer, the overhead cost originally protected by the threshold will be turned to profit without protection and dissipated by competition.²

No matter what will the factor change cause, or no matter whether market competition protects the overhead cost of an industry or dissipates the profit (turned from previous overhead cost) of it, it must be good to the economy as a whole. Because even if the profit of insiders is dissipated, it is different from the rent dissipation due to infringement of PPRs, but benefits consumers through price reduction, which means the previous overhead cost of insiders is converted into consumer surplus, a concept to be introduced in the later lecture "Price-searching". Thus, from the perspective of the whole economy, market competition does not dissipate overhead cost, but converts them into consumer surplus. Market competition still protects the rent of the whole economy in the way that the rent of the whole economy now is not in the form of producer surplus (overhead cost), but consumer surplus.

Fourthly, the size of overhead cost also affects the quality of product, which is determined by "the Principle of Rent Matching". It is a theory with wide application and strong explanatory power, which will be introduced in detail in the next section.

1. In MSE, the shutdown decision is analyzed with the wrong concepts of short-run, long-run and fixed cost. [↩](#)
2. The factor change (such as technological progress) leading to decrease entry threshold may be bad to insiders, but may also be good. A decline in monopoly in an industry will lead to an increase in the supply, then resulting in a decline in price, and demand will rise according to the law of demand. If the scale of the whole industry remains unchanged or just increases little, insiders must suffer from the decreasing overhead cost. However, if the demand rises dramatically, so the scale of the whole industry grows significantly, the income of insiders will increase instead. The development of automobile industry after China's entry to the WTO is a typical example. Therefore, the factor change such as technological progress or liberalization of entry control which will lower the threshold or monopoly is not necessary to dissipate the overhead cost of insiders. The final effect depends on the change in the overhead cost of the whole industry. [↩](#)

Section 4: The Principle of Rent Matching

The principle of rent matching can be expressed by a Chinese saying "a good horse should be matched with a good saddle". A good saddle is a saddle of high quality (rent), while a good horse is also a horse of high quality (rent). A horse of high rent should be matched with a saddle of high rent. Although the ancient Chinese did not know economics, they could discover this economic principle through a large number of observations of facts. They knew the fact, but did not know the reason.

So, what is the reason why a good horse should be matched with a good saddle? It can be understood from several perspectives.

Firstly, the principle of rent matching is one of the implications of the law of demand.

If one has bought a high-grade residential land with picturesque scenery, he would not build a bunk house there. If he did, the construction cost could not be recovered when he sold the house, and only the price of the land could be recovered. A house built on a good residential land must be also good, which is the implication of the law of demand.

A good residential land is a land of high rent, which must be matched with a house of high rent. Why? Suppose the price of a land is RMB10 million, while the construction cost of a bunkhouse is RMB1,000, which means that the price of land per unit construction cost for the bunkhouse is RMB10,000. However, if a luxurious house with a construction cost of RMB1 million is built on the land, the price of land per unit construction cost for it is RMB10. The law of demand says that when price (cost) falls, quantity demanded will rise. Therefore, self-interested people will choose to build a luxurious house of RMB1 million instead of a bunkhouse of RMB1,000.

There are other logically similar examples. In a community, those houses in the middle or other good locations are all large ones with high prices, while those that are small and with low prices are all located at the edge of the community or other poor location. The reason is the same. On the land with the highest rent in a community, developers will choose to build the houses of the best quality which not only refers to the quality of construction and design, but also includes the living quality with the least noise from the outside of the community. By contrast, on the land with the lowest rent, developers will build the houses of the worst quality.

Secondly, the principle of rent matching can be understood from the perspective of recovering overhead cost.

A reporter has interviewed a flowerpot seller at an auction for booths in Flower Market for Spring Festival, and asked what kind of flowerpot he would sell. He replied, "I don't know until the auction is over. If the auction price is high, I will sell high-quality flowerpots, and vice versa." This seller is just like ancient Chinese mentioned above who did not know economics, but he can discover this principle through experience that the flowerpots of high-quality (rent) should match the booth of high auction price (rent). But why? Because the rent of booth is overhead cost, and the high overhead cost needs the goods of high price to recover, and only the high-quality flowerpots can be sold at a high price.

Suppose the rent of a booth is RMB1,000, if a bottle of water with the cost of RMB1 is sold, it must be priced RMB1,001 to cover all the cost (including the cost of water and the booth). Will anyone buy such expensive water? Obviously no! By contrast, if a diamond with the cost of RMB10,000 is sold, it must be priced RMB11,000 to cover all the cost. Will anyone buy such an expensive diamond? Of course yes!

The rent of production factors like shop is overhead cost that producers need to obtain income from the market as much as possible to recover. The higher the rent, the more they need to sell high-quality products of high price to recover.

There are other logically similar examples. It is high-quality goods that are sold in the prime locations. Luxury goods are always sold in luxury shops in the deluxe business circles of big cities, and will never be sold on stalls in the rural countries or the back lands.

Thirdly, the principle of rent matching can be also understood from the perspective of equal marginal principle.

According to equal marginal principle, if a good horse matches a poor saddle, the marginal quality of the poor saddle will be higher due to the effect of the law of diminishing marginal returns which will be discussed in the later lecture "Market Structure: Price-taking". Caution: the marginal quality of a poor saddle is high, but the total quality of it is low. Therefore, when the total quality of the saddle is improved, the marginal quality of it will decline, so that the marginal quality of the good horse and that of the saddle tend to be equal, closer to satisfying equal marginal principle, thus improving the overall quality of the horse with saddle. If the quality of the saddle is continually improved, the law of diminishing marginal returns will continue to work to make the marginal quality of the saddle continually decline, eventually fall to be the same as the marginal quality of the horse. Thus, equal marginal principle is satisfied, and the overall quality of the horse with saddle reaches the maximum. The quality can be regarded as rent, so the total rent of the horse with saddle reaches the maximum when the marginal rent of the saddle is equal to that of the horse. From this perspective, it is clear that to what extent the rent is optimal.

In management, there is a so-called theory of "Cask Principle" or "Cask Effect" which says just as the content of a cask slab, not the longest one, the competitiveness of a firm is determined by the weakest link, not the strongest one. Strictly speaking, it is not a theory at all, but a sum up of some similar phenomena, which is precisely an implication of the principle of rent matching from the perspective of equal marginal principle.

The principle of rent matching implies that it is more effective to improve the competitiveness of the weakest link of a firm than to improve that of the strongest one. Since the competitiveness of the weakest link is weak, the marginal competitiveness of it should be higher than that of other relatively stronger links. Therefore, according to equal marginal principle, if the competitiveness of the weakest link is improved to be equal to that of the other relatively stronger links, the overall competitiveness of the entire firm will be improved too. The competitiveness of the weakest link means the rent of it is low, while the competitiveness of strongest link means the rent is high. The combination of low rent and high rent is out of line with the principle of rent matching, so to improve the competitiveness of the weakest link is to increase the rent of it to match the high rent of the other relatively stronger links, which result in the rise of the total rent (overall competitiveness) of the whole firm.

Some may ask why not sell low-quality flowerpots in a high-rent booth to earn more income through smaller margins with greater volume so as to recover overhead cost? The answer is: it depends on whether the price elasticity of demand for low-quality flowerpots is elastic, which means a small fall in the price can bring about a substantial rise in quantity demanded, so the income finally increases. The problem of elasticity has been discussed in Lecture 8 that in economics, until now there are no reliable theories to infer the elasticity of a good in advance. There is another similar but opposite question: why not sell high-quality products in a low-rent booth? Is it not easier to recover overhead cost than selling in a high-rent booth? The answer is: who dares to buy a high-priced product in a low-rented booth? For example, who dares to buy a diamond from stall? No one, because buyers will not believe it is a real diamond. The key constraint is information cost. Low-quality products are low-priced, and even if they are fakes, the buyers will not lose much money.

If there is no problem of information cost, such as the buyers and sellers in the school canteen are familiar with each other, so the information cost is almost reduced to 0, the poor-decorated canteen will also sell high-quality goods. Here, what matches the high-quality (high-rent) goods is actually the good-credit instead of the poor-decoration of the school canteen.

Thus, when the principle of rent matching principle is applied to explain phenomena, we should be careful. There are some phenomena that seem to have refuted this principle. For example, goods sold in tourist attractions are often expensive and of poor quality, as in service areas of expressways. The rent of these locations is high just because of administrative monopoly which is different from the monopoly caused by the market competition such as the prime locations in cities. Take the service area of expressways which is far away from residential districts as an example. The passengers have to buy food there, or else they will wonder how far to go before they can reach a residential district. By contrast, the prime locations in cities are easily subject to competitive pressure from the locations nearby. If the goods sold there are only high-priced without high-quality, consumers will leave there and choose to do shopping nearby.

In brief, if there is administrative monopoly, the effect of the principle of rent matching will be distorted. After all, this principle is based on the pressure of market competition.

Section 5: The Form of Contract Affects Overhead Cost

As mentioned before, the calculation formula of overhead cost is “income - direct cost”. In the previous sections, only the change of overhead cost caused by the change (decline) of income has been analyzed. However, the factors that affect overhead cost are not only income, but also the form of contract which is regarded as the other unchanged factors before. In this section, income is regarded as the other unchanged factors, and the effect of different contract forms on overhead cost will be analyzed.

We will take the business of selling vegetables as an example to show that even if income is unchanged, the overhead cost under different contract forms will be very different. Although the business of selling vegetables is a commercial (trade) activity, the relevant analysis can be analogized to industrial production.

Suppose one is thinking about the business of selling vegetables. After investigating the relevant income and cost, he finds that the retail price of vegetables in the market is RMB1, while the wholesale price 0.2RMB which is equivalent to the cost of raw materials in industrial production. He needs a carrying-pole to transport the vegetables to the market, and the rent of it is 0.2RMB a day which is equivalent to the cost of machines and equipments in industrial production. In the market, he has to rent a stall to sell the vegetables, and the rent of it is 0.3RMB a day which is equivalent to the cost of the factory buildings in industrial production. He also needs to hire a labor to sell the vegetables, and the wage is 0.2RMB a day which is equivalent to the labor cost in industrial production. The money he needs is borrowed from a bank, and the interest is 0.1RMB a day.

Before entering the market, all the costs have not been paid, so all the costs are direct costs. For simplicity, suppose one only sells one unit of vegetables a day. All the costs are RMB1, which are equal to the income (the price of the vegetables). Therefore, one decides to enter the market according to the zero-profit theorem.

One borrowed RMB1 from a bank, and paid the interest of 0.1RMB immediately, thus the interest cost became historical cost. And he hired a worker and paid him the wage of RMB0.2 a day immediately, thus the labor cost also became historical cost. He then instructed the worker to rent a carrying-pole, and paid the rent of RMB0.2 immediately, thus the rent of it became historical cost. The worker went to the wholesale market and spent RMB0.2 to buy a unit of vegetables, thus the cost of the vegetables also became historical cost. The worker transported the vegetables to the market with the carrying-pole, and rented a stall and paid the rent of RMB0.3, thus the rent of the stall also became historical cost. In this way, all expenditures (RMB1) are historical costs and there are not any direct costs. Therefore, as long as the market price of the vegetable is not 0, even if it is lower than the expectation of RMB1, one should accept the price, and all the income is overhead cost.

However, one can obtain all these production factors in another way. When borrowing money from a bank, one agreed with the bank not to pay fixed interest, but to share the income with it, which is equivalent to the bank taking a stake in the capital and earning 10% of the income from selling vegetables, so he did not have to pay a fixed amount of interest immediately. Then he agreed with the worker that the wage was not a fixed amount but a share of 20% of the income from selling vegetables. When renting the carrying-pole, the rent was not a fixed mount, but a share of 20%. When buying the vegetables in the wholesale market, he agreed with the wholesaler that the price of the vegetables was a share of 20%. The worker carried the vegetables

to the market, and leased the stall with a rent of a share of 30%. So far, one has not paid a cent, so there was no historical cost, and all is direct cost.

In other words, if all the production factors are obtained by sharecropping contract, all the costs is direct cost and there is no overhead cost. By contrast, if all the production factors are obtained by one-time buyout or fixed-rent contract, there is no direct cost and all the costs is overhead cost.

In reality, all kinds of production factors are usually obtained through a mixture of fixed-amount contract and sharecropping contract. Generally speaking, fixed-amount contract is mostly used for capital, vegetables and carrying-pole, while labor is usually paid with base salary (fixed salary) and commission or bonus (sharecropping wages), and location rent is often a combination of fixed rent and sharecropping rent.

The nature of the costs in fixed-amount contract is closer to overhead cost, while the cost in sharecropping contract must be direct cost. The nature of the former cannot be completely sure, because even if it is a fixed-amount contract, unless it is a one-time buyout, the longer the contract term, the closer it is to the nature of overhead cost, whereas it is closer to direct cost.

For example, the wage of hourly workers is closest to direct cost. The daily wage is overhead cost within one day and becomes direct cost beyond one day. By analogy, the weekly, monthly and annually pay is overhead cost within one week, one month and one year respectively, and becomes direct cost beyond one week, one month and one year respectively. Even if the term of a contract is long, if the contract is allowed to be terminated halfway, overhead cost can become direct cost. Of course, it depends on the compensation required to terminate the contract. If the compensation is rather high, which means that the transaction cost is high, it is closer to the nature of overhead cost. However, if income falls sharply, even if the transaction cost of legally terminating the contract is very high, producers still have the option of illegally breaking the contract.

Usually, producers with higher overhead cost are more able to bear the decline in income than those with lower overhead cost. As mentioned before, the income from the market is the (total) cost. When income falls, the producers with lower overhead cost will leave the market earlier, because the decrease of income has exhausted overhead cost and cuts into direct cost, which means the producers are eliminated by competition. Those who have only direct cost and no overhead cost are called "marginal producers" because they are on the edge of the industry. As soon as the income falls, which cut into the direct cost, they have to leave the market immediately. However, if all the production factors are obtained by through sharecropping contract, although all the costs is direct cost which can fall along with the income, the decease of income will never cut into direct cost! In other words, those who obtain all the production factors through sharecropping contract and all the costs is direct cost are the most likely to survive in an economic recession.

However cautious: if the production factors obtained through sharecropping contract can earn higher income elsewhere, they also have the option to terminate the sharecropping contract at any time, which will cause the producers to leave the market.

For example, suppose the daily wage is 20% of the income from selling vegetables, the worker can only earn RMB0.1 when the price of vegetable falls to RMB0.5. If he can find another job elsewhere with a daily wage of more than RMB0.1 (no matter it is a fixed wage or a sharecropping one), obviously he will choose to terminate the sharecropping contract and job-hops.

In other words, if the income decline is not due to the overall recession of the industry or economy, but only due to the poor performance of the boss, even if the sharecropping contract is used, the income distributed to the production factors will be lower than that they can earn elsewhere, so they will still choose to terminate the sharecropping contract, which causes the producer to be eliminated out of the market. The reason why they can earn higher income

elsewhere is, of course, that the lower income here is simply due to the poor performance of the producer. Thus the function of "survival of the fittest" in the market competition will still work even with sharecropping contract.

In any case, since wages can be flexibly and automatically adjusted downwards along with income decline, there can be no unemployment in a world where all production factors are obtained through sharecropping contract. The boss will not fire the employees, but it is possible that the employees will fire the boss. Then why are all production factors not obtained through sharecropping contract? This is, of course, because the transaction cost of sharecropping contract under certain constraints are too high. In other words, that there is unemployment implies that there is transaction cost, which is the reason of unemployment. It will be discussed in detail in the later lecture about unemployment.

In a word, when the income (i.e. the cost) is unchanged, the structure of cost, which means how much overhead cost is and how much direct cost is, depends on the form of contract through which the production factors are obtained, and the transaction cost of legally or illegally terminating the contract. The next lecture on transaction cost will explain that the reason why there are different forms of contract is also due to the different transaction cost. Therefore, that how much is overhead cost depends not only on income, but also on the specific transaction cost under different constraints.

Thus, it is clear what the fundamental mistake of the theory of firm in MSE is. It is impossible to know the structure of cost without considering transaction cost. And if the structure of cost is not clear, it is impossible to infer even the most basic behavior of producers such as shutdown or not. Marshall used the concepts of short-run and long-run to avoid the fact that there is transaction cost in reality, which is a serious mistake.

Lecture 14: Transaction Cost

Section 1: Transaction Cost is Indispensable

The concept of transaction cost has been mentioned as early as Lecture 3, but it has not been formally explained until now. This concept is not easy, so although it is so indispensable, it is not suitable to formally introduce it as soon as it comes up. Because transaction cost is a kind of cost, it is more appropriate to introduce it after the concept of cost has been explained. In addition, this concept should be introduced before the lecture about monopoly, because it is impossible to understand various behaviors of a monopolist without the constraint of transaction costs. That is why this lecture about transaction cost is put after that of cost and before that of market structure.

Some may think that the textbooks of MSE can only introduce the relatively simpler phenomena where there is no transaction cost. So it is a kind of step by step learning, and not a mistake. However, if the textbooks of MSE really only talks about the topics that do not involve transaction cost, there is very little to talk about. In fact, they again and again step into the field where the phenomena cannot be analyzed correctly without transaction cost.

For example, if transaction cost is not taken into account, it is impossible to explain the postulate of self-interest or rationality correctly. That one is self-interested or rational does not mean that he will not make mistakes. Under the constraint of information cost (which is a kind of transaction cost), a rational person may make a wrong choice. If transaction cost is not taken into account, it will seem that he is not rational by choosing to harm himself. That is why the MSE contradict itself by saying that self-interest or rationality is the postulate of economics and saying that there are irrational behaviors at the same time.

In addition, whether price criterion or non-price criteria should be used to determine the result of competition also depends on the constraint of transaction cost. That normative economics thinks there are inefficient behaviors stems from ignoring transaction cost. The concept of disequilibrium is also due to ignoring transaction cost. Government intervention in price (such as price control) does not result in shortage or surplus but the rise in transaction cost (rent dissipation is also a kind of transaction cost). The theory of firm can only be an engineering theory without considering transaction cost. Different forms of contract (caused by different transaction costs) lead to different structure of cost. All these topics cannot be analyzed correctly without transaction cost. Can they be excluded from the textbooks of MSE? If yes, what else is left?

The concept of transaction cost does not easy, but it is so important to economics that is it indispensable, just as air to human. Avoiding it will only lead to many mistakes and contradictions in the textbooks of MSE.

Section 2: The History of Transaction Cost

What is transaction cost? For this complicated and difficult concept, it is necessary to know the history and follow the whole process of the invention and development of this concept.

Coase was the first economist to invent this concept. He published an article entitled "the Nature of Firm" in 1937, and put forward the concept of "transaction cost", which was one of his achievements in winning the 1991 Nobel Economics Prize.

Why did Coase invent the concept of transaction cost? He wanted to answer a question in that article: why are there firms? As mentioned at the beginning of Lecture 11, the traditional theory of firm only cares about the problem of input-output which is about how much output a certain input can produce, focusing on the technical characteristics of production, and does not care about the production organization or contract arrangement. In other words, it is essentially an engineering theory, not an economic one at all. No matter in principle or in practice, production does not necessarily have to be organized through firms. Producers can be self-employed individuals or firms. However, in reality, especially in modern society, a large number of productions do be organized through firms. So the question arises: why are there firms? The market can also organize productions. Why are some productions organized by firms?

Some economists believe that teamwork is more efficient than individuals working independently, which is why firms are needed. However, teamwork or cooperative production can also be organized through the market. In reality, many self-employed individuals buy semi-finished products from upstream producers and process them into those closer to finished products, and then sell them to downstream producers for further processing until finished products can be directly sold to consumers. These individuals are linked to teamwork, but they are organized through the market instead of by the boss or manager in a firm. Of course, there is also much teamwork organized through firms. Thus the question remains: why are there firms? The market can also organize cooperative productions, so why are some of them organized by firms?

Coase's answer is: because there is cost for market to organize productions, which is transaction cost.

Coase did not give a clear definition of transaction cost, but listed the cases where there is transaction cost.

For example, a self-employed individual must firstly buy inputs (production factors) before production, so he must find upstream producers who can provide the inputs to him. It will take time, energy and money which are all transaction costs (information costs). After finding the proper upstream producer, he has to bargain the price, quality, quantity, etc. and may need to sign a contract, which also take time, energy, and money. After the contract is signed, it needs to be executed. He must supervise whether the supplier delivers the goods on time with good quality and quantity in accordance with the terms of the contract, which also causes transaction cost. If the supplier breaks the contract, this individual producer has to negotiate with him. If the negotiation fails, he may have to go to the court and hire a lawyer to file a lawsuit, so the transaction cost will be even higher.

After buying all inputs, this individual produces the outputs by himself. Now he needs to sell the outputs, which means he has to find sellers. He may depend on the distributors to sell his products to consumers indirectly, and there will be additional transaction cost. Even if the products are sold directly to consumers, there will also be all kinds of transaction cost in finding

consumers, bargaining prices, quality and quantity with them, signing contracts, executing contracts, negotiating and even filing lawsuits when contracts are broken.

The market is the same as an ordinary commodity in that there is cost for using it. As Friedman said, there's no such thing as a free lunch, so the market is neither free. In Coase's view, it is the cost of using the market to conduct transactions, so he calls it transaction cost.

Coase pointed out that the benefit of using firms to organize production is to avoid using the market and thus to avoid paying transaction cost. If the producers in upstream and downstream are just different departments in the same firm, the department in downstream does not need to spend time, energy and monetary to find the upstream producers who can provide inputs, but directly asks the department in upstream to provide inputs. Neither does it need to bargain the price, quality and quantity, but just needs to transport the inputs from the warehouse of the department in upstream to the workshop of the department in downstream according to the instruction of the boss or manager of the firm. And it does not need to sign any contract or even worry that the contract may be broken.

One more example. Suppose one works as a secretary in a firm. In principle, secretarial service can be bought through the market. Imagine that every day one goes to that firm to see what secretarial service the manager needs today, such as how many pages of documents to type or copy, how many cups of tea to serve to visitors, etc. For simplicity, suppose there is only the work of copying. One bargains with the manager to determine the pages of documents, the required quality of copying, the price for the copying service. It is the same as that the manager takes the documents to the copy shop and pay for the service. Now one is not an employee of the firm, but a self-employed individual who runs a copy shop that provides on-site service.

Obviously, it is too troublesome! Every day one goes to the firm without knowing if there is any need for copying, how many it needs, and when it needs, which are all information cost. He also needs to bargain with the manager about the quantity, quality and price of copying the documents, and may even sign a contract. That it is troublesome means there is very high transaction cost in economic terms.

What is more, secretarial service is actually much more than copying, and includes typing, serving tea, and welcoming visitors. Does one need to bargain with the manager about the price and quality for each service, such as how hot the tea should be, how much water should be in a cup, how fast the tea should be served to the visitors, how one should smile when serving the tea ... The transaction cost of using the market is too high. Therefore, we seldom observe the phenomenon of trading secretarial services through the market. More generally, firms recruit secretaries in the labor market, roughly describe the work of secretary, and set a price according to the time (time wage). When one is hired as a secretary in a firm, he must obey the instructions of the manager, without bargaining! That is precisely why using firms can save transaction cost: because one can no longer be choosy and bargain as can be done in the market.

In other words, the market guides individuals to provide products or services through prices: only when they think the price is appropriate, will they provide goods. By contrast, in a firm, they are ordered to provide products or services through the instructions of the boss or manager: no matter they like it or not, as long as they are employees of the firm, they must obey the orders. It seems that transaction in the market is free, while the work in a firm is compulsory. However, there is still a market outside the firm. One is free to choose which firm he likes to work in. If he thinks that the salary the firm offers to him is not worth what he contributes by obeying the manager's order, he can ask for a salary raise. If he cannot, he is free to quit and find another job. Similarly, if the manager thinks one's performance is not worth his salary, he can ask for a pay cut. If he cannot, he is free to fire him and hire someone else.

As mentioned before, some economists think that teamwork or cooperative production is the

cause of firm. In fact, if cooperative production is organized through the market, the transaction cost is generally very high.

There are two cases. One is the cooperative production along successive links in a chain, which needs to be closely linked in time. For example, every step in a production line can be organized through the market in principle, but the products need to move continuously from upstream to downstream in the production line like running water, so there is not much time for the individuals in each step to bargain.

The other case is that several production links need to cooperate simultaneously. For example, when shooting a film or a TV drama, the cooperation of scenery, lighting, clothing, photography, actors, etc. can be organized through the market in principle, but all need to be carried out at the same time, so they cannot bargain with each other.

Because there is transaction cost, when using the market, one has to go through the process of finding counterparties and bargaining the price, quality and other relevant contract terms, which is rather time-consuming for cooperative production. Therefore, all these production links are put into a firm, and cooperate by obeying the orders of the boss or manager.

Section 3: From the Special Transaction Cost to the General Transaction Cost — Methodology of Science (4): Ad Hoc Theory

There is a problem in Coase's explanation of the cause of firm by transaction cost of the market. Since using firms can avoid transaction cost of the market, why are all productions not organized through firms? Obviously, if there is only cost in using the market, the postulate of self-interest implies that all productions should be organized through firms, and the market should completely disappear. However, in reality, there is still market everywhere. Is it disequilibrium? As pointed out in lecture 7, the seemingly disequilibrium implies that there must be constraint overlooked.

The answer is obvious: there must be also cost for using firms! As mentioned in the previous section, firms instead of the market are used to organize productions, which means the boss or manager's orders instead of the market price are used to guide the various production factors.

As pointed out in Lecture 6 about the cause of the Great Depression and Lecture 9 about the theory of supply and demand, as long as there is no problem with monetary policy, price is an indicator that transmits information about supply and demand, quality and so on to all parties in the society quickly and generally accurately. People do not need to know what the specific demand and supply are in the market, just adjust their behaviors along with the change in price. The behaviors of different people guided by price are all adjusted in the direction of reaching equilibrium. By contrast, within a firm, there is not price to guide the options of different production factors which only blindly obey the manager's instructions. The manager can only roughly estimate based on internal information feedback and his experience, so the chance of making mistakes is higher than the price as the guideline.¹

What is more, as pointed out in Lecture 4, the market system is price criterion which is a kind of competition criterion. Thus the market is replaced with firms, which means the price criterion is replaced with non-price criterion, and it will lead to rent dissipation. What exactly causes rent dissipation in firms? Among all kinds of production factors, labor is very special. Labor is more effectively motivated by price than any manager's instructions in the world.

Neoclassical economists are used to dividing production factors into two categories: labor and capital. In reality, there are hundreds of production factors, why are they divided into these two categories? Neoclassical economists may say because raw materials, machinery, equipment, land and factory buildings ... are all obtained with money, so they are classified as capital. However, labor is also obtained (rented) with money and why is it not a kind of capital? In fact, labor does be a kind of capital because it can be called human capital.

The reason is: influenced by Marshall, neoclassical economists like to use geometry as a mathematical tool, especially rectangular coordinate system. There are two axes for a rectangular coordinate system, so there are two variables needed. In consumer theory which is used to analyze product market, the two axes represent the quantity of two goods respectively. In the theory of firm, neoclassical economists similarly used two production factors to analyze factor market. However, just as the two goods in the product market only need to refer to good X and good Y, the two production factors in factor market do not need to be specifically designated as labor and capital, and only need to refer to factor X and factor Y.

In the traditional theory of firm which ignores transaction cost and the influence of organizational form or contractual arrangement on production, there is no essential difference between labor and other non-labor factors. However, if the influence of organizational form is taken into account, there will be significant difference between labor and non-labor, because human has his own will, and will not do what he is asked to do. A machine will automatically and consciously operate even without supervision. Human is different. Without supervision, one may try to shirk. Therefore, the boss must send another one to supervise him. But who supervises the supervisor?

There are two kinds of supervision, which is the same as a farmer driving a donkey. A carrot can be hung in front of the donkey to attract it to move forward, and a stick can be used to beat the donkey when it refuses to move forward. A carrot refers to award as incentive, while a stick refers to punishment. If it is a tractor rather than a donkey, obviously the farmer does not need carrot and stick. In other words, compared with non-labor factors, there is additional cost of carrot and stick for labor. If the production is not organized through firms, and the producer is a self-employed individual, naturally he does not need a carrot to motivate him or a stick to punish him, but the price of the market directly motivates and punishes him. Thus, only when there are two constraints (to use labor rather than non-labor in a firm rather in the market), there will be additional cost of carrot and stick.

Coase called the cost of using labor in a firm as management cost. However, the cost of using the market is called transaction cost, while the cost of using firm is called management cost, which is the mistake of "ad hoc theory" in methodology of science.

In Lecture 2, the mistake of tautology has been introduced, which is a statement with empty contents, and is impossible to be refuted. Ad hoc theory is the other extreme, which is a statement with too many contents, and is too easy to be refuted as soon as the constraints change slightly.

Let us take an example of natural science to illustrate. Suppose there are a mobile phone and a cup on the table. Both are stationary, because they are all subjected to a pair of forces with opposite directions and the same magnitude, and are in equilibrium. One is the force upward of the table supporting them, while the other is the force downward. What is the force downward? One can define it in this way: the force on the mobile phone is called "force of phone", while that on the cup is called "force of cup". This kind of definition is precisely a typical case of ad hoc theory. When the object to be explained changes from a mobile phone to a cup, the definition of "force of phone" is immediately refuted. Similarly, "force of cup" is also limited to only explaining a cup.

Is it logically similar that Coase defined the cost of using the market as "transaction cost" and the cost of using firms as "management cost"? These two concepts with different names can only explain the cost of the market and that of firms respectively. Once the constraint changes from the market to firms or vice versa, these two concepts will be refuted immediately. In social science (economics), the problem of ad hoc theory is not so easy to be observed as in natural science.

There is no difference in essence between transaction cost and management cost, and a more general concept should be used to unify them, just as "force of phone" and "force of cup" are essentially the same force which is the gravity of the earth. Similarly, in economics, the transaction cost of using the market and the management cost of using firms are essentially the same cost which is transaction cost of using different forms to organize production. Self-interested human will compare the transaction costs of the different organizational forms. With the same production efficiency, human will choose the organizational form with the least transaction cost. It is just like in the market, with the same quality, a buyer will choose the good with the lowest price.

Thus, after Coase, the concept of transaction cost is expanded to include the cost of using firms or management cost.

Let us continue to compare physics with economics. Newton's concept of gravity is actually ad hoc theory to some extent, because it is only applicable to explain the force of the earth on objects. In fact, there is force of objects on the earth, which is too small relative to the force of the earth. According to Newton's law of $F = ma$, the influence of the force of objects on the earth's motion (measured by acceleration) is insignificant. However, the mass of stars in space is no longer negligible with respect to the earth, and the concept of gravity should be further generalized to "universal gravitation" which is commonly among various objects.

Logically similarly, the concept of transaction cost in economics also needs to be further generalized to a concept in broader sense. It is Steven. N. S. Cheung who further expands the concept of transaction cost. The market is a kind of institution, and a firm is also a kind of institution. Besides the market and firms, there are other organizational forms in the world, such as family, organization and nation.

As mentioned in Lecture 4, the transaction cost of using the market is not only the cost of finding upstream producers, bargaining price and quality, and so on, but also the cost of game rules supporting the price criterion (i.e. the cost of the legal system protecting PPRs). It is the same for non-market system. For example, there are rules and disciplines in a firm to support the order of the boss or manager, and moral customs in a society, laws and regulations in a nation, family rules in a family, and even in gangdom, there are so-called rules and regulations. Different organizational forms are actually carriers of different game rules or institutions. It is those different institutions that specifically define different organizational forms and apply different competition criteria. They not only organize productions, but also restrict all kinds of behaviors in competition. The use of different organizational forms is actually the use of different institutions, which is not free lunch, and there is different cost for the establishment, operation and maintenance of different institutions. Human choose between different institutions, just as they choose between different goods, and of course, they will choose the institution with the lowest cost.

However, when the market is not used, there is no transaction, so the concept of transaction cost sounds misleading literally. Thus, the so-called transaction cost should be more generally called Institution Cost which is the cost of the establishment, operation and maintenance of institution. Because the concept of transaction cost initiated by Coase has become too popular, it can still refer to institution cost. So it is the compromise: the cost of using the market is the special transaction cost, while the cost of using different institutions is the general transaction cost which is institution cost.

Thus, the function of all kinds of institutions (including the market) is to decrease some kinds of transaction cost, but any institution will inevitably increase other kinds of transaction cost. A certain institution can survive only because the transaction cost it increases is lower than the transaction cost it decreases. For example, if human choose to use the market under a certain constraints, it must be because the transaction cost of using non-market (such as firms, planned economy, etc.) is higher. Although the market will also cause some kinds of transaction cost, it must be relatively lower than other kinds of transaction cost the market can decrease.

Of course, the change in constraints will lead to the change in the transaction cost of different institutions, which may cause human to switch from one institution to another. It is similar to the change in the relative price of different goods, which may cause consumers to switch from one good to another. This is precisely institutional transformation.

By comparing the process of generalizing the concept from "gravity" to "universal gravitation" in physics with that of generalizing the concept from "special transaction cost" to "general transaction cost" (institution cost) in economics, we can not only better understand why the scope of transaction cost is continuously expanding, but also better understand how useful theories (i.e. scientifically truth) are constructed. Ad hoc theory is too easy to be refuted, while tautology is completely irrefutable. Therefore, although they are opposite in nature, the results are the same in that they are neither useful theories. Ad hoc theory and tautology are two extremes on a line, and the useful theories should be located between them by adding appropriate constraints.

There are two ways to construct a useful theory by moving to the middle from the two extremes. One way is to start from tautology and move toward ad hoc theory by adding constraints, which is the basic application of marginal analysis introduced in Lecture 7. A typical example is that the economics is constructed based on the postulate of self-interest (a tautology) with different constraints.

The other way is to start from ad hoc theory and move toward tautology by relaxing assumptions, which is called "generalization". Here, the development of the concept of transaction cost is a typical example. Firstly, starting from the specific phenomenon of why there are firms, Coase put forward the concepts of transaction cost and management cost, which is ad hoc theory. Other economists followed up the development and gradually generalized to the concept of institution cost which can be used to explain all kinds of organizations.

In scientific methodology, there has always been a dispute between "deduction" and "induction". In fact, they are only these two ways to construct useful theories. Start from tautology is based on unreal postulate, and seems to be purely logical deduction. However, if constraints as objective facts are not added, it will be unscientific because it is irrefutable. On the other hand, starting from ad hoc theory is based on objective facts (experiences), and the commonness (laws) of a large number of similar phenomena are sought to realize generalization. Both roads lead to Rome, and they are just applied to different phenomena.

1. F. A. Hayek (1899-1992) was the first economist to expound the role of price in reducing the information cost for both suppliers and demanders, and in reaching equilibrium in his article "The Use of Knowledge in Society". What he called "knowledge" is information mentioned in this book, so this article is probably the earliest study on information cost in economics, and even earlier than the concept of transaction cost put forward by Coase. [←](#)

Section 4: The Perspective of Contract

Steven. N. S. Cheung generalized the concept of transaction cost to institution cost. However, in order to avoid being too general and becoming tautology without refutability, he advocated to observing transaction cost from a more specific perspective. Different organizational forms are carriers of different institutions, but what on earth are institutions? As mentioned in the previous section, they are the contracts in the market, the rules and disciplines in a firm, the moral customs in a society, the laws and regulations in a nation, the family rules in a family, and the rules and regulations in gangdom.

In fact, they are all contracts! It can be a written contract or a conventional consensus. The contracts signed between parties in the market are certainly those human are used to regarding as contracts, but the rules, disciplines, moral customs, laws and regulations seem to be far from contracts. When one is employed by a firm, he has signed not only a contract of employment, but also accepted the regulations and disciplines made by the firm which is a common contract accepted by all employers of the firm. There are also implicit common contracts (such as practice and national laws) in the market. In western ideology, national laws are "social contracts" signed between people and government. Moral custom is said to be a contract between human and God, because in western myth, the rainbow in the sky after rain is the sign of a contract between human and God.

Of course, the latter two seem to be mandatory compared with the former two. In the market, one can choose not to trade and reject the other party's contract, while in a firm, one can choose to resign and refuse the rules and disciplines of the firm. However, in a society or a nation, it seems difficult for one to choose not to follow the moral customs or abide by the law. In fact, there are always choices. One can choose to leave the nation he does not like, or live an independent life in a wild mountain region. Just the cost is too high, so most will not choose to do so. It is similar to the cost of suffering a loss or even losing the job if one rejects the other party's contract in the market or of the firm. It is also similar to one's obstinate refusal to buy anything because he thinks all goods in the market are too expensive and of poor quality, and then he may starve to death.

From the perspective of contract, the different organizational forms or different institutional arrangements are actually different contractual arrangements, or specifically, different contractual terms. The contract terms which directly measures and sets price are based on price criterion without rent dissipation (a kind of transaction cost), are called price terms or income terms.

However, the cost of measurement (also a kind of transaction cost) is not zero. When it is too high, human will choose to measure indirectly by other contract terms (use terms). Indirect measurement will make the measured person tend to perform the contract ostensibly, causing rent dissipation which is an inevitable result of non-price criteria. Thus, although the cost of indirect measurement is relatively lower, there will be relatively higher supervision cost (management cost or transaction costs). Human will trade-off between higher cost of direct measurement and higher supervision cost of indirect measurement, seeking the lowest total transaction cost of a combination of contract terms, which is similar for consumers to trade-off between different goods, seeking the lowest total cost of a combination of goods.

Thus, what Coase said about replacing the market with firms is actually replacing another combination of contracts (of product sales) with one combination of contracts (of factor purchase or lease). When human choose different institutions, they actually choose different contracts, and the institutional transformation is the change of contractual terms.

Until now, the concept of transaction cost (institution cost) can be introduced. Steven. N. S. Cheung put forward the most appropriate definition of transaction cost as the cost that does not exist in Robinson's one-man world. He further defined "society" as a world with more than one person. Once there is more than one person, there will be competition between them due to the postulate of self-interest and the constraint of scarcity, and there will be institution cost (transaction cost) because the competition criteria need relevant institutions to support their establishments. In other words, transaction cost is the cost of coordinating the competitive relationship between people and is the cost of society.

Some may ask, "Is the cost of pregnancy transaction cost?" or "Is the cost of a game that needs at least two players such as tennis transaction cost?" Of course not. It is possible for technology to make artificial embryo, sperm or ovum, and even make human single-breeding. It has been possible to simulate an opponent in playing games. Any problem that can be solved by natural science is not related to transaction cost. There are problems of transaction cost that cannot be solved by natural science in the world is precisely why economics or social science is needed besides natural science. The theory of firm with production function as the core in traditional economics does not take transaction cost into account, so is just an engineering theory in essence, not an economic theory at all. Similarly, economics without the concept of transaction cost cannot be real economics.

By the way, what is the relationship between transaction cost and the direct cost and overhead cost mentioned in the previous Lecture 13? There are both transaction cost belonging to direct cost and that belonging to overhead cost. After the establishment of an institution, the transaction cost incurred during the establishment has become historical cost and the recover of it should be treated as overhead cost. When the institution is operating and maintained, the relevant transaction cost is direct cost. For example, the commission paid to the intermediary of a transaction according to the price is a typical transaction cost belonging to direct cost, because the role of the intermediary is to assist the buyer and the seller to find each other and reach a transaction agreement, and what he earns is the transaction cost.

Lecture 15: Market Structure – Price-Taking

After the introduction of the theory about cost, next it is the market faced by producers should be analyzed.

In MSE, product market is separated from factor market. In product market, producers are the suppliers, while individuals or households are the demanders. By contrast, in factor market, producers are the demanders, while individuals or households are the suppliers. However, as will be discussed in the later lecture about production factors, if the transaction cost is low enough and the production is organized by piece-rate contract, product market cannot be separated from factor market. The market structure to be discussed in this and the next lecture is applicable to any markets, and product market is taken as an example here.

Section 1: The Definition of Price-Taking

As for market structure, in MSE, it is divided into two categories: "pure competition" (perfect competition) and "monopoly". The former is also called "Price-Taking", because producers can only passively take the price prevailing in the market. The latter is also called "Price-Making" because producers can indirectly make price by changing outputs.

However, we advocate that the former should only be called "Price-Taking" rather than "pure competition" or "perfect competition", while the latter should be called "Price-Searching" rather than "monopoly", and the term "Price-Taking" is used to describe another behavior of making use of information cost to increase price.

Why?

Firstly, the term "pure competition" or "perfect competition" is misleading. As mentioned in Lecture 4, competition is everywhere in a society which is not Robinson's one-man world. So even under the so-called monopoly market structure, there is still competition, just the form of which is different from that under the price-taking market structure. The problem is what "pure" or "perfect" means. No matter what market structure is, any competition must be restricted by the legal system protecting PPRs, and not completely free, so the term "pure" or "perfect" is meaningless.

Secondly, the concept of monopoly in MSE is seriously wrong, which will be the theme of the next lecture. Why the term "price-making" should be replaced by "price-searching" will be also explained in detail in the next lecture.

In this lecture, we will introduce the market structure of price-taking first. What is price-taking? It is most appropriate to define it by the shape of the demand curve. When the demand curve faced by a single producer is a flat line¹, it is the market structure of price-taking, because he can only take the price where the demand curve intersects with the vertical axis representing the price. If he set the price higher than it, consumers will turn to other producers, so he cannot sell out a single unit. And if he set the price lower than it, he suffers a loss from the perspective of opportunity cost, because he can sell out numerous units at the market price which is the opportunity cost of selling at other prices). In other words, no producers can set price higher than the market price and will set price lower than the market price, so they actually have no right in setting price and can only take the market price.

The market structure of price-taking is actually very rare, because it appears only when the following two conditions are satisfied:

(1) All producers supply the same products among which there is no difference in type, quality, or even location. Even if there is slight difference, the price will be different. If two stores sell exactly the same products, but one is located in a residential area and the other is far away from the residents, the former can charge a higher price, which is the rent of the location, making the same product more valuable.

(2) There is no information cost. When information cost is high, the producer can benefit from his knowing more than the consumers by price-making, which is due to the rent of information.

Why is the demand curve faced by a single producer in a price-taking market a flat line? The explanation of MSE is that in the so-called pure competition, there are countless producers and consumers, so a single producer accounts for only a tiny share of the whole market. In other words, what a single producer faces is only a small point on a demand curve of the whole market

sloping downwards. If enlarged, that point seems almost like a flat line.

However, there is a serious mistake in this explanation in that the quantity of producers and consumers in the market is actually irrelevant. As will be pointed out in the next lecture, it is a serious mistake for MSE to define monopoly as “there is only one producer in the market”. The quantity of producers in the market has nothing to do with whether the market structure is price-taking or price-searching (monopoly). Even if there is only one producer in the market, it can be price-taking. And even if there are numerous producers in the market, it is completely possible to be price-searching as long as there is high information cost. Not only those who are in the market and be observable are competitors, but also those who are outside the market but are eager to come in are potential competitors. Therefore, what matters is not the quantity of producers in the market, which will be discussed in detail in the next lecture.

1. Note: it is not the demand curve of the market which is always a curve sloping downwards. [←](#)

Section 2: The Optimum of Producers

When the postulate of self-interest is applied to the theory of firm, the problem of equilibrium or optimum is how a producer chooses the appropriate outputs to realize rent maximization.

Mathematically, it is to seek the extreme value of $R - C$ (the difference of revenue and cost), which is to find the derivative of it to be 0 (i.e. $(R - C)' = 0$ or $R' = C'$). Derivative is a marginal magnitude in economics, so this formula can be rewritten as $MR = MC$, which is precisely the optimum for producers. It is universally applicable, not only to price-taking discussed in this lecture, but also to the price-searching to be discussed in the next lecture. It is actually a specific form of the third extended application of marginal analysis (equal marginal principle) mentioned in Lecture 7.

The above is from the perspective of mathematics to explain the optimum of producers, while the following is the economic implication.

For a producer, if he produces and sells one more unit of output, on the one hand he can obtain more revenue (marginal revenue or MR), but on the other hand he has to pay more cost (marginal cost or MC). If $MR > MC$, it means the production of this unit of output is worthy, because it increases the total rent. Thus he will choose to produce one more unit of output, which implies that he has not reached the optimal output, and should continue to expand the production scale.

However, if he produces and sells one more unit of output and finds $MR < MC$, it means the production of this unit of output is a loss, because it decreases the total rent. Thus he will not produce more, but may consider if he should cut one unit of output instead. If he does so, on the one hand his revenue will fall, but on the other hand he can also save some cost. Because now $MR < MC$, it is worthy to cut one unit of output, which increases the total rent. Thus he will choose to cut one unit of output, which implies that he has produced too much and exceeded the optimal output, and should reduce the production scale.

In a word, when $MR > MC$, the output is less than optimum. And when $MR < MC$, the output is more than optimum. Only when $MR = MC$, can the output be optimum, because now a more (or less) unit of output leads to the same additional income and additional cost, which means the total rent will neither increase nor decrease, and actually has reached the maximum.

Specifically in the market structure of price-taking, since a producer can only passively take the prices prevailing in the market and the market price will not change no matter how many outputs he sells, the average revenue (AR) and marginal revenue (MR) are both equal to the market price (P), which is shown in Fig.15-1 where AR, MR and the demand curve (d) that the producer faces are the same flat line.

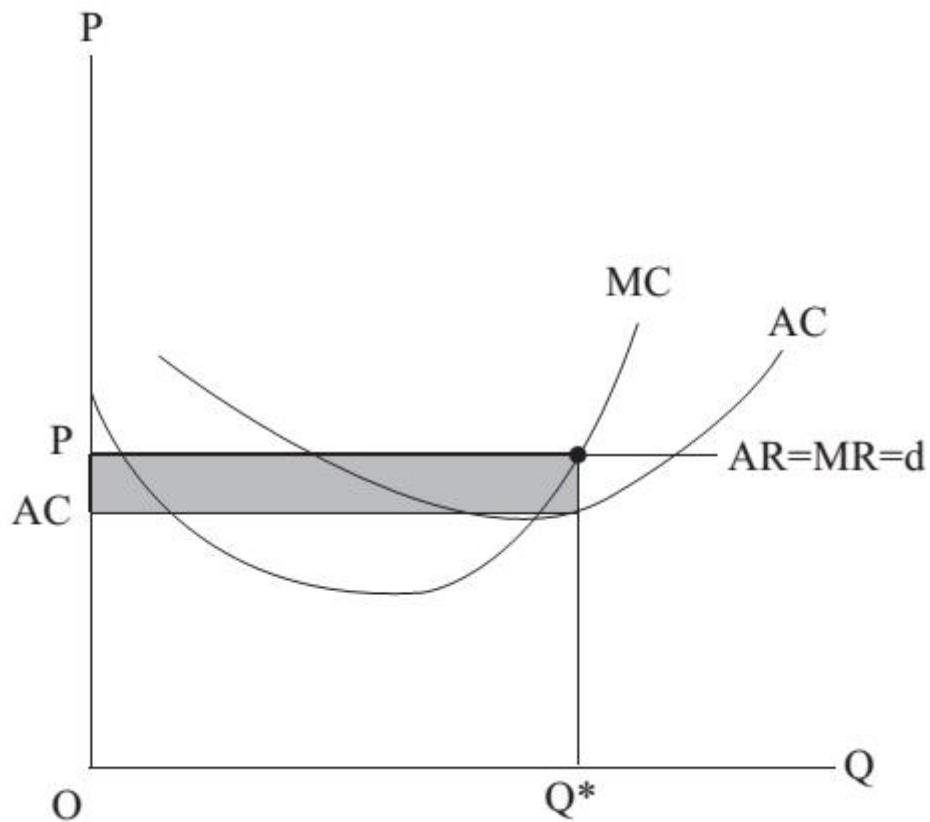


Figure 15-1

Section 3: The Law of Diminishing Marginal Product

As for cost, there are also average cost (AC) and marginal cost (MC). As shown in Fig.15-1, AC is a u-shaped curve, while MC is also u-shaped that passes through the bottom of the AC from below. In MSE, there is a u-shaped curve called average variable cost (AVC) which is shown in Fig.15-2 and is located below the AC and the bottom is also passed through by the MC.

For these three curves of cost in MSE, there are some questions.

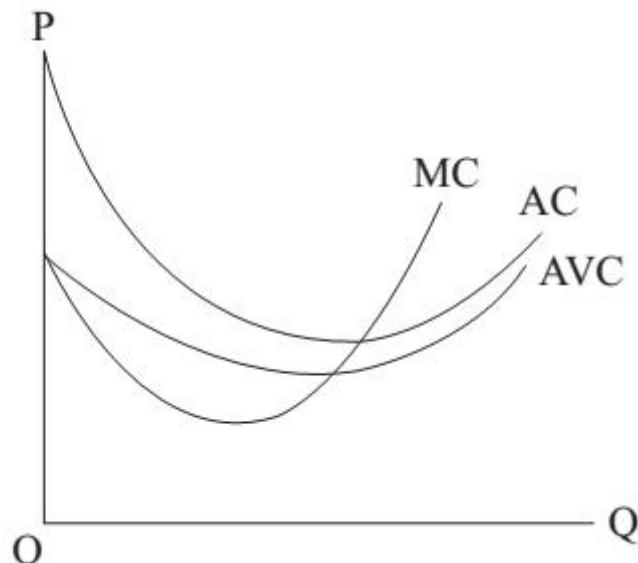


Figure 15-2

(1) Why is MC curve U-shaped or why does MC decrease first and then increase?

The answer is "the Law of Diminishing Marginal Product", also known as "the Law of Diminishing Return". It is an iron law, which means it must be true. This law says, as long as one of the production factors is fixed, as the inputs of other factors increase, the total output will increase at the beginning, but finally it will turn to decrease. Because the increase in total output (i.e. marginal product) will gradually decrease, eventually total output will turn to decrease when marginal product has already decreased to a negative value.

The law of diminishing marginal product can be proved by reduction to absurdity. If this law does not hold, people can grow enough food for the entire world in a flowerpot. The size of the flowerpot holds constant the amount of soil which is one of the production factors. Continuous watering and fertilization in the flowerpot can initially increase the total output, but after a certain point, the total output will increase more and more slowly and eventually turn to decrease. Biology tells us that too much water will drown the plant and too much fertilizer will burn the plant to death. While economics tells us that there is an optimal ratio of various production factors. Once the optimal ratio is exceeded, on the one hand the redundant will not increase output, and on the other hand they may even hinder production and lead to a drop in output. Therefore, even if all production factors increase, but one of them does not increase proportionally or increases more slowly than the others, the law of diminishing return will also work. If there were no the restriction of this law, producers would choose to expand production endlessly, which has never been observed in reality, so this law is an iron law.

Strictly speaking, the law of diminishing product is not an economic theory, but a natural law which is also applicable to natural science. However, the constraint of transaction cost can be added to make it significantly important in economics. Before, the constraint of transaction cost is added to amend the traditional theories in MSE. While this law of diminishing marginal product has never been refuted by facts, and the complementation of the constraint of transaction cost further strengthens its conclusion.

There is an ancient Chinese saying “one monk carries water by himself, two monks cooperate to carry water, and three monks have no water”. In the production of “carrying water”, the production factors are labor (monks), two buckets and a pole. When there is only one monk, he uses a pole to carry two buckets by himself. Due to his limited strength, the water in the buckets is only half full, and he can only walk back and forth for once. When there are two monks (the input of labors increases), they can change the mode of production, from one person carrying the water by himself to two persons cooperating to carry the water, so they can fill the buckets with more water and walk back and forth for more times and increase the total output. In other words, the increase of input of production factors enables producers to choose some more efficient production mode (technology), thus increasing output.

However, when there are three monks, the additional monk will do no good to increasing output. With only one pole and two buckets, three persons can not cooperate to carry more water than two persons, but only hinder each other. If two of them carry the water first, and then the third one cooperates with one of them to carry the water next time, there is always one monk who is idle and does no good to output increase. That is why even the total output will increase, but the increase (i.e. marginal product) will be less than that of the input of labor increasing from one monk to two monks, which is precisely the implication of the law of diminishing of marginal product.

Taking the implication of this Chinese saying into account, one more monk will make it more difficult to cooperate with each other, which means the transaction cost will rise. When the input of labor from two monks to three monks, the transaction cost of cooperation increases to such an extent that no one is willing to carry water, and the total output falls to 0! The traditional law of diminishing marginal only considers factors of natural science, but if transaction cost is also taken into account, the effect of it will be further strengthened.

When the decreasing part of the marginal product curve is multiplied by the price of product, it will become the demand curve of production factors. Since no producer is so stupid to stop expanding production at the increasing part of the marginal product curve, the meaningful demand curve for production factors must be sloping downwards, which is not conflict with the law of demand.

Suppose the production factor is labor (L) and the price of it is W , then

$$MC = \frac{dC}{dq} = \frac{WdL}{dq} = W \bullet \frac{dL}{dq} = \frac{W}{\frac{dq}{dL}} = \frac{W}{ML}$$

Thus, the shape of the marginal product curve is directly related to that of the marginal cost curve, except that the former is inverted U-shaped, while the latter is U-shaped, and the law of diminishing marginal product translates into the law of rising marginal cost.

(2) Why is AVC curve U-shaped?

In MSE, there are production costs of short-run and long-run. In short-run, there are fixed cost and variable cost, while in long-run there is not any fixed cost. AVC here is an average magnitude by dividing variable cost (VC) by output. As has been criticized in Lecture13, fixed cost in MSE is a wrong concept because it is in essence historical cost which is not (opportunity) cost in economics, and should be replaced by the concept of overhead cost. While the concept of

variable cost (VC) should be changed to direct cost (DC), so AVC should be also changed to Average Direct Costs (ADC), and there should not be difference of short-run and long-run.

According to the law of diminishing marginal product, and adding the constraint of transaction cost, ADC curve will be U-shaped, and MC curve will pass it through the bottom, which has been shown in Fig7.2 of Lecture7.

Section 4: Average Cost

(3) Is AC curve U-shaped?

Since in MSE, there are production costs of short-run and long-run, there are also AC curves of short-run and long-run accordingly, which is, of course, wrong. Even if there were long-run AC curve, it could not be U-shape, but a flat line. Why? Because the so-called long-run means that no production factor is fixed, or the law of diminishing marginal product does not apply now, but the so-called "Returns to Scale" is applicable. In MSE, there is no law about returns to scale, and three cases are all possible: diminishing returns to scale, constant return to scale, and increasing returns to scale.¹ However, Friedman pointed out incisively in his "Price Theory" that with the assumption of equal proportion change of all production factors, logically only the constant returns to scale is possible and the other two cases are both impossible.

Here is Friedman's analysis. Suppose Paris's subway system is expanded as twice as before, can the passenger capacity be also doubled? The intuitive answer is impossible, and the passenger capacity should increase less than twice, because the expansion of the subway system will make the underground in Paris more crowded and there will be not enough demand from the population and business activities. In other words, it seems that the diminishing returns to scale will apply to this question. However, it is only because the underground space, population and business activities in Paris have not been expanded as twice as before, which means not all production factors increase by equal proportion, and it is conflict with the assumption of returns to scale. Just imagine all production factors really increase by equal proportion (even the air is doubled), it is equivalent to clone another Paris, and the passenger capacity of the subway will of course be as twice as before. Thus, only the constant returns to scale is possible, which is always true, and in fact a tautology.

However, if we get rid of the wrong division of short-run and long-run in MSE, producers can only determine ADC and MC. There is no marginal magnitude for overhead cost, because there is overhead cost even without production. There is only one kind of AC, regardless of short-run or long-run, but it is determined by the market price. The income of a producer obtained from the market should be deducted direct cost, and the rest is overhead cost. The income divided by output is the market price, while the total cost (the sum of direct cost and overhead cost) divided by output is AC, so AC is equal to the market price. In other words, the demand curve faced by a producer is precisely his AC curve! Thus, AC curve is a flat line in the market structure of price-taking, and a line sloping downwards in the market structure of price-searching, and both are not U-shaped. However, on AC curve, what is really meaningful is only the point corresponding to the optimal output (satisfying the equilibrium condition of $MR = MC$), and it is not necessary to draw the whole AC curve.

The AC curve is always the same as the demand curve, which means $P = AC$. There may be a profit more than the expected income ex post, which is not opportunity cost. In fact, taking into account all opportunity costs (whether it is overhead cost as rent or transaction cost), there can not be the so-called "profit" or "supernormal profit" calculated by $P - AC$ in MSE, which is the implication of the "zero-profit theorem". In MSE, this theorem says, "In the long run, (economic) profit for any firm in a competitive industry is zero." There are two major mistakes in it. Firstly, there is no need for the wrong division of "long run". Secondly, there is no need for the restriction of "competitive industry" (i.e. the market structure is price-taking). The correct zero-profit theorem should be: Profit for any producer in any market (whether it is price-taking or price-searching or so-called monopoly) is zero. In MSE, it is believed that there is supernormal profit for the so-called

monopolist, which in fact is monopoly rent. Rent is in essence cost, not profit. It will be explained in detail in the next lecture.

The implication of zero-profit theorem is that in reality, there is the same so-called (accounting) profit rate for all producers in all kinds of industries. Suppose there are only two industries A and B in an economy. When one engages in industry A, he has given up the income from engaging in industry B, so the profit of B is the opportunity cost of engaging in A. If his profit is lower than that in B, he has suffered a loss, and he will turn to B. If his profit is higher than that in B, the producers in B have suffered a loss, and they will turn to A. In other words, the profit for all producers in all kinds of industries must be the same, or else producers will compete to leave the lower-profit industry and enter the higher-profit one.

In fact, this same profit rate is the rate of return on investment (ROI), which appears as the market interest rate for loaning. People will compare the market interest rate with the profit or rate of ROI in their own industries. If the profit is lower than the interest rate, they would better shutdown their business and deposit the money in the bank. If the profit is higher than the interest rate, they will borrow the money from the bank to do business for this supernormal profit.

Some may argue, there is obviously higher profit in some industries or firms. There can be three reasons. Firstly, it may be the rent that is mistaken for profit. Some have unique vision or shrewd tactics, so their income is far higher than that of the ordinary. It is the return (rent) on their human assets of vision or tactics. Secondly, it may be due to administrative monopoly. The supernormal profit is still rent, but it is monopoly rent, which will be explained in the next lecture. Thirdly, it may be an industry with high risk, so the seemingly extraordinarily high profit is actually a "Risk Premium" to compensate those who bear high risk.

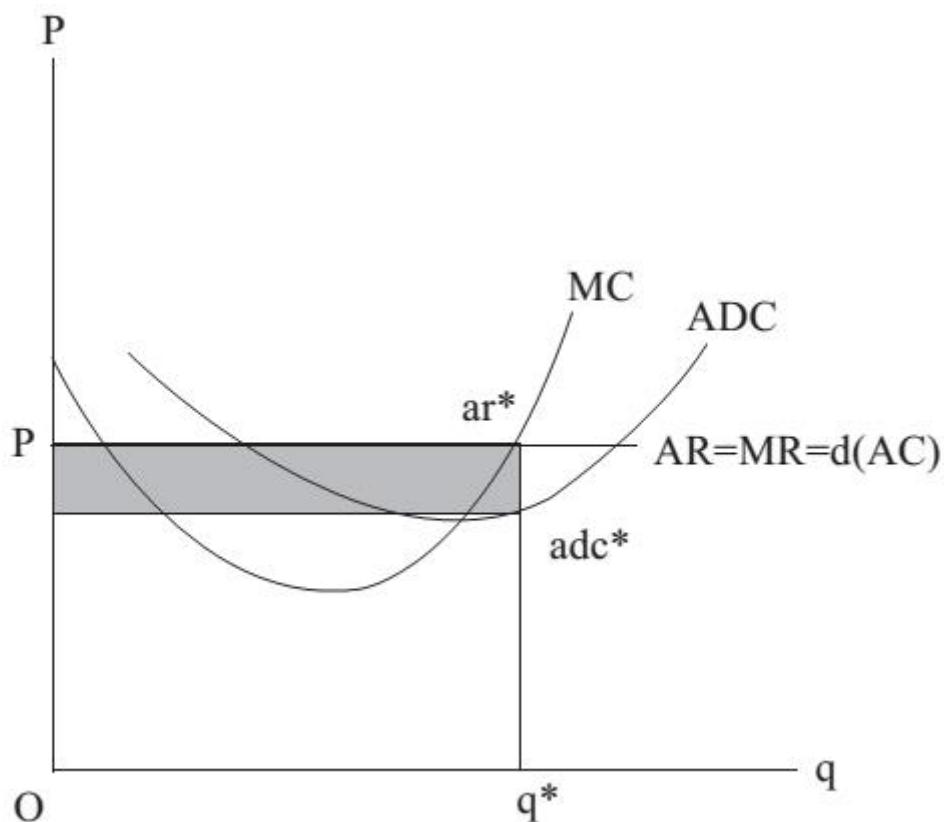


Figure 15-3

Venture Capital (VC) is the most typical example. A successful VC may have tens of thousands of times of returns, but the vast majority of VC will lose all the investment. This financial innovation, originally born in Silicon Valley, was once regarded "profiteering" and was heavily taxed by the Ford and Carter Administration of US, which led to a sharp drop in the amount of VC during those

years. Fortunately, it was not long before Reagan became the President and reduced the tax rate a lot.

According to the above analysis, there is following mistake in the previous Fig.5-1: the AC curve should be ADC curve (AVC curve in MSE) as shown in Fig.15-3. The correct AC should be the flat line that coincides with AR, MR and the demand curve, but what is meaningful is only the point corresponding to q^* (i.e. ar^*). According to $MR = MC$, the producer's optimal output is q^* determined by the intersection of MC and MR. The point of adc^* corresponding to this optimal output is located below the point of ar^* , so the shaded area of $(ar^*-adc^*) \times q^*$ is the rent or overhead cost.

1. Cautious: do not confuse the "diminishing returns to scale" with "the law of diminishing returns". The former applies to the so-called long-run case which means all production factors are variable, while the latter applies to the so-called short-run case which means there is at least one of production factors is fixed. [e](#)

Section 5: The Theory of Supply and Demand with Production

Next we will explain how to deduce the theory of supply and demand with production based on the producer's optimum. What has been deduced in Lecture 9 is the TSD without production.

Firstly, as shown in Fig.15-3, the demand curve and AR curve are the same line, which is correct no matter in the market structure of price-taking or price-searching, as long as a producer sets a single price for all products. Because $AR = \frac{R}{q}$, when a producer sets a single price for all products, $R = P \times q$ which is put in the previous formula, and get $AR = P$. If it is the market structure of price-searching (monopoly), the demand curve sloping downwards reflects how Q changes along with the change in P, while the AR curve reflects how AR (like P) changes along with the change in Q, so they appear to be the same.

Secondly, as shown in Fig.15-4, the part of MC curve above its intersection with ADC curve is the supply curve of a producer. Why? Because a producer will always choose the optimal output according to $MR = MC$, the equilibrium point formed by the combination of optimal output (q^*) and price (P) must move along the MC curve with the change in P, which is precisely the same as the supply curve reflecting how Q changes along with the change in P. However, as has pointed out in Lecture 13, once the income from the market falls below the direct cost, a producer will choose to shut down the business, and the output (supply) will drop to 0. Thus, there will be no supply or supply curve below the bottom of ADC curve.

Section 6: Consumer Surplus and Producer Surplus

If the demand curve and supply curve are drawn in the same diagram, it represents the theory of supply and demand already introduced in Lecture 9.

Now it can be explained more deeply. What does the demand curve in Fig.15-4 actually represent? A point on the demand curve represents the use value of that unit of good for the consumer, and is also the highest price he is willing to pay for it. Remind the concept of use value in Lecture 10, one is willing to pay a price for a good because it is useful and valuable to him. How much is one willing to pay for the good? The answer is: he is willing to pay a price not higher than the use value of the good for him. The use value of a good will decrease when the quantity of it increases, which is called "the Law of Diminishing Marginal Utility" in MSE. As mentioned in Lecture 10, the concept of utility should be replaced with use value, so this law should be also renamed as "the Law of Diminishing Marginal Use Value", which seems similar to "the Law of Diminishing Marginal Product". It is true that they are logically symmetrical, except that the law of diminishing marginal product is an iron law faced by producers, while the law of diminishing marginal value is used to describe consumers' behaviors.¹

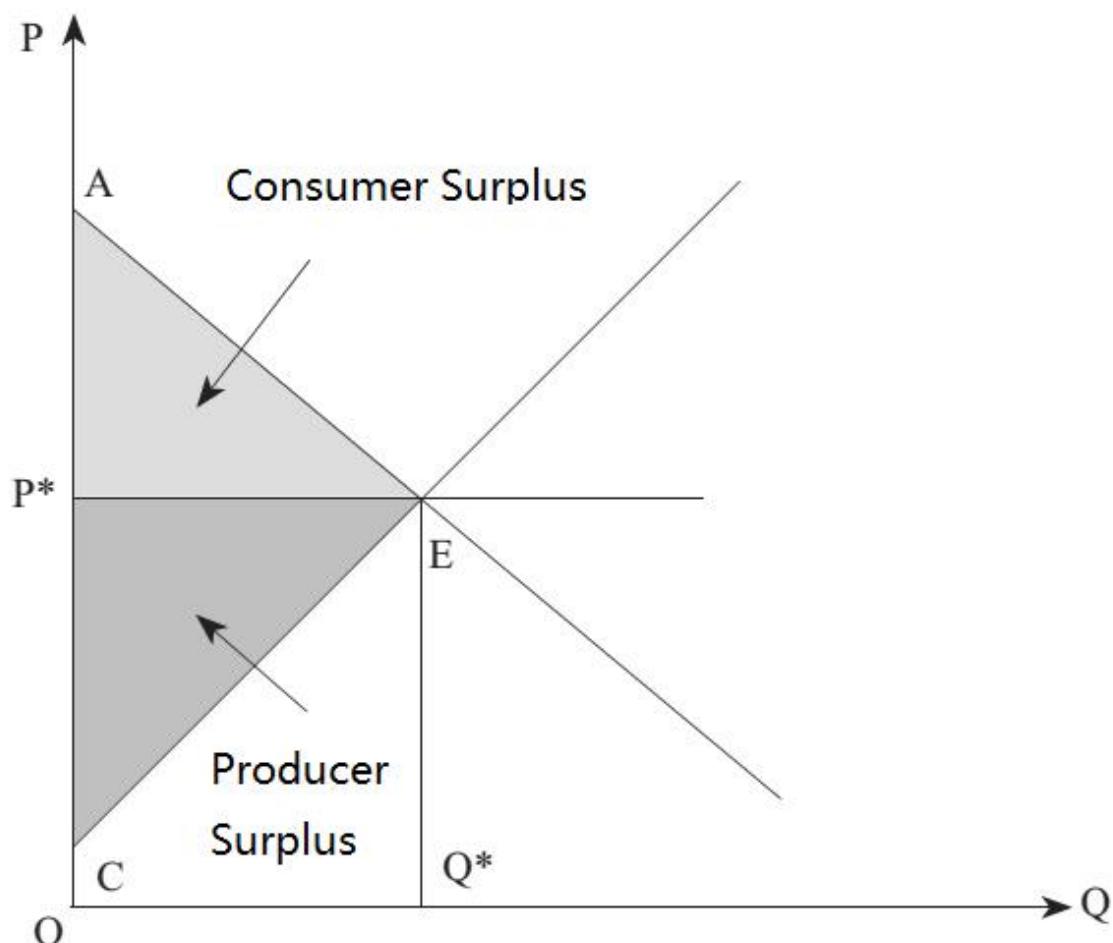


Figure 15-4

The law of diminishing marginal use value can be proved by the following example: when one is hungry and starving, and there is a piece of bread, how much is he willing to pay for it? Obviously, even if he must give up everything he has to buy the bread, he is willing to do so, because the use value of the bread is equivalent to his life. Without the bread, he will starve to death, while without his life it is meaningless to have much money.

However, when he has had this piece of bread, and there is another piece of bread, how much is he willing to pay for it? Because he has already had the previous piece of bread and is not as hungry as before, he is not willing to pay as much money as before.

When there is a third piece of bread, he is willing to pay further less for it. The use value of each new piece of bread offered to him will become lower and lower. After a certain point, he will even have too much bread and no longer feel that he is enjoying delicious food, but that he is fed up, which means the use value of a more piece of bread has fallen to negative for him. It is precisely why the demand curve is sloping downwards, which is actually due to the law of diminishing marginal use value.

In the market structure of price-taking, no matter how much a consumer buys, the market price is the same. The postulate of self-interest implies that the optimum for a consumer is to buy until the use value of the last unit is equal to the market price (i.e. point E in Fig.15-4). If one buys a unit more than E, the use value of this last unit of good or the marginal use value is less than the market price value which is the cost of this last unit of good or the marginal cost. In other words, the optimum or equilibrium for consumer is “marginal value = marginal cost” ($MV = MC$). For a consumer, the use value of a good is the non-monetary revenue from consuming the good, so the above equation can be generalized as “marginal revenue = marginal cost”, which is precisely the same as the optimum or equilibrium for producer. As mentioned in Section 2, $MR = MC$ (equal marginal principle) is universally applicable.

Before the equilibrium point E, the use value of each unit purchased by a consumer is higher than the market price P he pays, so there is surplus for the consumer. If producer charges a higher price, the consumer is still willing to pay because before the last unit of good, the highest price he is willing to pay is higher than the market price. In the next lecture some pricing arrangement for producer to charge higher price will be introduced. Compared with those cases, there does be surplus for consumer in price-taking. It is called “Consumer Surplus”. Sum up all consumer surplus of each unit of good, it is a triangle-like shadow enclosed by the vertical axis above the price and below the demand curve (if the demand curve is a curve, it is a curved triangle).

The above analysis can be similarly applied to producers. The supply curve of a good represents the lowest price a producer is willing to charge. The supply curve is sloping upwards, which represents the increasing marginal cost, because it is part of the MC curve. Once a producer enters the market, he will continue to produce until the marginal cost increases to be equal to the market price (point E in Fig.15-4) because before that point the market price (MR) is higher than the cost of producing a more unit of good (MC).

In the market structure of price-taking, before point E, the revenue of each unit of good (P) is higher than MC, so there is surplus for producers, which is called “Producer Surplus”. Attention: the cost here is direct cost, because only for direct cost there is marginal magnitude. Therefore, producer surplus is actually overhead cost! Sum up all producer surplus or overhead cost of each unit of good, it is a triangle-like shadow enclosed by the vertical axis below the price and above the supply curve (if the supply curve is a curve, it is a curved triangle).

In MSE, consumer surplus and producer surplus are used to explain that transaction is beneficial to both buyers and sellers (consumer surplus for buyers and producer surplus for sellers), and are used as tools to measure welfare. Therefore, they are widely used in welfare economics to judge government policies.

However, it is not really the importance of these two concepts. Producer surplus is actually overhead cost, the importance of which has been fully explained in Lecture 13, while the importance of consumer surplus will be fully explained in the next lecture about price-searching.

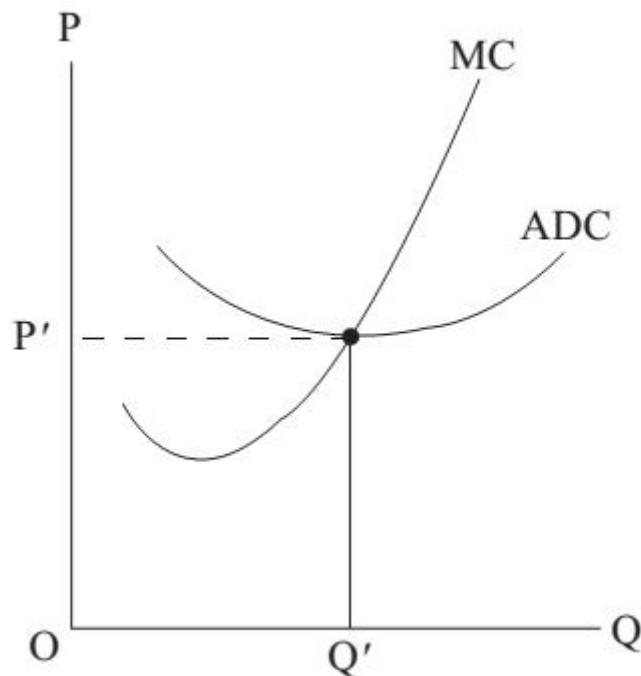


Figure 15-5

In Lecture 9, it has pointed out that the supply curve and the demand curve are actually the same, but what was analyzed then is the case without production, while now is the case with production. Is the above conclusion that the supply curve is a part of the MC curve conflict with the statement that the supply curve and the demand curve are the same? The answer is No! When there is no production, from the perspective of opportunity cost, the price of holding a good is to give up the alternative of selling it and holding the other good instead. Therefore, when the price of a good rises, if one continues to hold it instead of selling it, the cost of doing so rises on the margin, which is the marginal cost. In other words, even without production, a consumer's demand curve for a certain good is his MC curve for the other good, and MC curve is supply curve.

The above is the supply curve of a single producer, and the derivation of the supply curve of the whole market is the horizontal summation of the supply curves of all producers in the market, which is the same as that of the demand curve. However, there is a complexity here that if only a producer increases his output, it is just like a drop of water falling in the sea, and there is no influence on the market. If a large number of producers or even all producers increase their outputs, it is completely different.

Firstly, the increase in supply (of products) in the product market will lead to an oversupply and a drop in price if there is no change in demand, which will make producers be not willing to increase supply.

Secondly, there will be a large increase in demand for production factors in the factor market, which may lead to an increase in price unless the relevant factors are also widely used in other industries, and the impact of the increase in demand in this industry is negligible. The price of production factors is the cost of producers, which means that the marginal cost will rise more than when only one producer increases his output.

In other words, when the supply curves of all producers are summed up horizontally, the marginal cost curve (supply curve) of a single producer will actually change, resulting in the supply curve of the market summed up horizontally by them also changing. Based on the analysis above, it should be steeper than before, which means producers are willing to supply more products only with higher price due to the rise in marginal cost. However, it is only one of the possibilities. The supply curve may also become flatter, which means the marginal cost of a single producer decreases instead. The former case is called "external diseconomy", while the latter one is called

"external economy". The word "economy" here refers to cost savings that is the original meaning of it.

As has introduced in section 3, the curve of ADC (average direct cost) is U-shaped, which means when the output increases, the ADC of a producer decreases first and then increases. The former decreasing part of it is precisely "internal economy", while the latter increasing part is "internal diseconomy". The word "internal" here refers to the decrease (economy) or increase (diseconomy) of the ADC that is caused by the increase of the output (scale) of the producer (within the firm). In other words, the internal economy and the internal diseconomy represent that the ADC curve of the producer falls first and then rises.

By contrast, the external economy and the external diseconomy do not occur within a firm, but are caused by the expanding scale of the industry outside a firm, which may be that other producers have increased their outputs, or new producers have entered the industry, and has also led to the decrease (economy) or increase (diseconomy) of the ADC of the producer. In other words, external economy and external diseconomy represent that the whole ADC curve of the producer moves up or down.²

How there is external diseconomy has been explained above by that the increase in output of all producers will lead to an increase in the price of production factors and so marginal cost. Next we will explain how there is external economy.

In China there are many specialized markets where there are a large number of producers of the same industry, which is called "Cluster". Such a cluster of producers will make the competition extremely fierce. A buyer is easy to know the market condition by walking around in the specialized market, while a producer is difficult to take advantage of the information cost. Why would producers still like to cluster in the specialized market? It must be because the benefits producers can obtain from cluster far outweigh the disadvantages, or the external economy outweighs the external diseconomy.

What external economy does cluster can lead to?

Firstly, cluster can save information cost. The cluster of the products of the same quality can reduce the cost for consumers to judge the quality difference of them. In addition, it can also save advertising expenses. In the vicinity of my hometown, there is a well-known 10-kilometer-long road lined with all kinds of furniture shops. Although there are many famous shops, there are more little-known shops. If a small shop opens alone at a place where is far away from other furniture shops, it must spend on advertising to publicize the location of it and attract customers to visit. The cost of advertising or publicity is information cost. But if it opens on the furniture road, those who need to buy a large number of furniture will go straight there, and naturally see it there, so it can save the information cost of publicizing it.

Secondly, cluster can also benefit from the cost reduction brought about by specialization. Let us take the furniture road as an example. There are transport fleets independent of furniture shops to deliver furniture for customers, which save the cost for furniture shops. If a furniture shop opens alone at a place far away from the furniture road, it has to keep a fleet to deliver furniture to customers. If customers are required to transport furniture by themselves, and it is difficult to find a suitable vehicle, they will tend not to buy furniture there or will not buy large size of furniture.

There will be independent transport fleets on the furniture road because there are a large number of furniture shops to supply a lot of business, which is enough to support independent fleets that rely solely on transportation service instead of selling furniture. With the large scale, the cost of an independent transport fleet is lower than that of a fleet kept by a furniture shop, because it is specialization for the former and has comparative advantage in cost. It is obviously due to the large scale of the entire industry, rather than a certain firm.

In addition, a furniture shop on the furniture road can be more specialized than that opens alone, which means it can only sell one kind of furniture such as desks because other shops can sell other kinds of furniture such as beds. Customers usually need to buy both desks and beds. If a furniture shop opens alone, it must sell all kinds of furniture to attract customers, which not only leads to higher cost, but also makes less rich varieties of each kind of furniture and reduces the attraction to customers.

Thirdly, cluster can also facilitate the producers in the same industry to cooperate with each other, thus reducing the transaction cost of cooperation. In China, firms are used to contract out some production to each other when they suddenly receive a large order with a tight delivery deadline. Obviously, if they are close to each other geographically, it is easier to cooperate in production which can be called "cluster in production".

There are many specialized wholesale markets (such as Yiwu small commodity market) in China, which are cooperation in sales. Regular customers often do not bother to go through such a large market to look for the goods they need, but go directly to a store with which they have established long-term and good relationship and tell it their needs. The store may have not some of goods, and the owner will help the customer to phone to the other stores for them.

The convenience of cooperation brought about by cluster is the key to the strong competitiveness of Chinese manufacturing in the international market. When a Chinese producer competes for an international order with the rivals from other countries, it does not fight alone, but is supported by the production capacity of the entire industry cluster in its region. How can foreign producers have such strong competitiveness to fight against the entire industry in China? In China, producers in the same industry will compete with each other, and will also cooperate with each other. Through competition, only the fittest can survive in the market, while through cooperation, the fittest will survive better in the international market.

1. Since "utility" in MSE is a measure of consumer preference, the law of diminishing marginal utility is regarded as a psychological law. [↩](#)

2. The increase in output within a firm leads to decrease first and then increase in the ADC (the point moves along the same ADC curve), while the increase in output of the whole industry outside the producer leads to decrease or increase in the ADC (the whole ADC curve moves downward or upward). It is logically the same as the quantity demanded change caused by price change (the point moves along the same demand curve) when other factors are unchanged, while the demand change caused by other factors change (the whole demand curve moves), which has been discussed in Lecture 8. [↩](#)

Lecture 16: Market Structure – Price-Searching

The market structure of price-taking analyzed in the previous lecture is rather rare in reality, while that of price-searching to be analyzed in this lecture is very common in reality.

Section 1: The Definition of Price-Searching

As mentioned in the previous lecture, in MSE, the market structure opposite to price-taking is usually called "monopoly" which is misleading. It is more appropriate to call it "price-searching".

Why is "monopoly" misleading? Because the definition of this market structure in MSE is wrong. In MSE, the market is believed to be monopolistic when there is only one producer in the market. The English "monopoly" comes from Greek, meaning "selling alone". However, when there is only one producer or supplier in the market, is it really monopolistic? As already mentioned in the previous lecture, the key feature of the so-called monopoly is that the producer (monopolist) has the pricing power, so monopoly is also called price-making. He can indirectly influence the market price by changing output, while the producers in the market structure of price-taking can only accept the price prevailing in the market. Therefore, the question is: when there is only one producer in the market, can he really have pricing power?

The answer is no. Suppose the market structure is price-taking, according to $MR = MC$, the optimal output q^* of a producer can be determined. With the equilibrium price, the total supply and demand in the market must be equal, which is supposed to be Q^* . And suppose there are n producers in the market and they are all the same, so $Q^* = nq^*$. Obviously, it is possible that $n = 1$ which means there is only one producer in the market in equilibrium. It has been supposed from the beginning that the market structure is price-taking, so when there is only one producer in the market, it is possible that he is not a monopolist.

How does it happen? Even if there is only one producer in the market, it does not mean that he can really have pricing power. If the market structure is actually price-taking, and the only producer naively thinks that he is a monopolist and raises the price, he will get an income higher than the opportunity cost or a "supernormal profit". According to the zero-profit theorem already introduced in the previous lecture, it cannot be equilibrium, because the potential competitors outside the market will try to enter the market to seize the supernormal profit.

For example, for the sake of simplicity, suppose that the demand in a market is only one unit of good, and the price that is in line with the market structure of price-taking is originally 10. If the only producer in the market raises the price to 15, there will be supernormal profit of 5. The outsiders will compete to enter the market by offering a lower price (such as 14). Remember that it is price-taking which means there is no difference in quality or location for different producers' supply, consumers will only choose the lowest priced goods, so the original producer will be eliminated for his higher price of 15.

However, the price of 14 is still higher than 10, and there is still supernormal profit of 4. The outsiders will continue to compete to enter the market by offering further lower price (such as 13) to eliminate the producer who offers 14. As long as the price is higher than 10, the only price in line with the market structure of price-taking, there will still be supernormal profit in the market and the zero-profit theorem is not yet satisfied. The outsiders will still compete to enter the market by offering lower price until the last directly offers the price of 10. Then there is no supernormal profit in the market any more, and neither will the outsiders have incentive to compete to enter the market.

The above is only an imaginary process. As long as the cost of entering the market (including transaction cost) is low enough, and the outsiders are smart enough (i.e. the information cost is low enough), the first producer who competes to enter the market will directly offer the price of 10 which is sufficient to eliminate the only producer in the market and will not be eliminated by other outsiders. In fact, if the only producer already in the market is smart enough, he will not be so foolish to believe he is a monopolist, and never dares to raise the price higher than 10, and he will know clearly that a supernormal profit does no good to him and will only attract potential competitors from outside the market.

A producer's competitors are far more than those visible in the market, but including huge potential competitors outside the market who will flock in at any time when there is supernormal profit in the market. It is shortsighted and stupid to think the only producer in the market is a monopolist with pricing power. Under some condition of cost, only one producer may be enough to meet all demands in a market, which does not mean the market structure is monopolistic and is possible to be price-taking.

Thus, the key to the market structure is not the number of producers in the market, but whether it is free to enter the market or, more precisely, whether the cost of entering the market is low enough. Let us continue to take the above example to illustrate. If it costs 3 for the outsiders to enter the market, the only producer already in the market can raise the price to 13. Although he can earn 3 more for each unit of good, the outsiders have to spend 3 first to enter the market, and they will suffer a loss if charge a price between 10 and 13 after entering the market. As a result, the only producer in the market can search a price that is most favorable to him at will between 10 and 13. It is precisely why the concept of monopoly should be revised to price-searching.

The entry cost (or, more vividly, the entry threshold) is the key to whether the market structure is price-taking or price-searching. With the protection of entry threshold, the producers already in the market (the number of producers needs not be one) more or less can influence the price by increasing or decreasing outputs to a level that is most favorable. The concept of price-searching or so-called monopoly should not be defined by the number of producers in the market, but by the shape of the demand curve faced by the producer in the market. When the demand curve faced by a producer is a flat line, the market structure is price-taking; while when the demand curve faced by a producer is a curve sloping downwards, the market structure is price-searching.

Someone may ask: in the market structure of price-searching, a producer can get more revenue (such as 3 more in the above example). Is it out of line with the zero-profit theorem? The answer is no. If there is no protection of entry threshold, how can the producer in the market search a price most favorable to him? The additional revenue of 3 is actually the return on the entry threshold, so it is the rent of it. In other words, the so-called "monopolistic profit" is not supernormal profit, but monopolistic rent due to the entry threshold. And rent is cost, not profit. A detailed analysis of how the entry threshold is formed can help us understand why the monopolistic rent is actually cost instead of profit. There are three kinds of formation.

(1) The entry threshold has been formed before a producer enters the market. The producer has to pay the entrance fee of 3 when entering the market. Before entering, it is direct cost. Once entered, it becomes historical cost which is no longer opportunity cost. If the expectation of the producer is correct, the difference of the income and direct cost (the overhead cost) should be 3, so the monopolistic rent of 3 only covers the payment of entrance fee, and not supernormal profit at all.

For the outsiders, before entering the market, the entrance fee of 3 is direct cost, so the direct cost of them is higher than that of the producer already in the market. Unless they can try to reduce the direct cost (in this example, it is 13), they cannot compete with him. And even if the direct cost of the outsiders falls, but does not fall below the direct cost of the producer already in

the market (in this example, it is 10), he can prevent them from entering by cutting the price into the overhead cost (in this example, it is 3). As mentioned in Lecture 13, the competition from outsiders will protect the overhead cost of insiders, while overhead cost will protect insiders from avoiding competition from outsiders. Therefore, the essence of the monopolistic rent or rent of the entry threshold is overhead cost.

(2) The entry threshold is built by a producer after entering the market. When the producer enters the market there is no entry threshold, so he has not paid any entrance fee. But after he has entered the market, in order to prevent the outsiders from competing with him, he builds an entry threshold of the market. Of course there is cost for the construction of the entry threshold, which is supposed to be 3 in the above example. Before the entry threshold has been built, the construction cost is direct cost, while after that, it also becomes historical cost.

If the expectation of the producer is correct, he can get higher income of 3 for each unit of good by price-searching, which will cover the construction cost of the entry threshold, so the essence of it is also overhead cost, not supernormal profit.

(3) The entry threshold is formed out of expectation. When the producer enters the market there is no entry threshold, so he has not paid any entrance fee. And after he has entered, neither has he built an entry threshold, so there is no construction cost. But an entry threshold is formed out of thin air due to the sudden change of some constraints after the producer has entered the market, so he is protected by it without paying any cost.

In this case, the additional revenue of 3 is an unexpected income or profit. However, once this unexpected luck has happened, the profit from the entry threshold will become opportunity cost or monopolistic rent, and it is no longer profit.

In conclusion, the value of the entry threshold for price searching or monopoly is monopolistic rent which is cost, or more precisely, overhead cost.

Section 2: The Cause of Monopoly (I): Innate Talent

There is another perspective to analyze how the entry threshold of monopoly or price-searching is formed.

(1) The cause of monopoly is innate talent given by God.

For example, the income of the sports stars or pop stars is far higher than that of the ordinary, which is a return on their talents in sports or music. This talent is unique to them, the outsiders cannot enter the market in any way, so it is a monopoly, and the return is monopolistic rent.

There are two key points for this kind of entry threshold.

Firstly, there may be no market value for a talent, so there is not necessarily monopolistic rent. Everyone in this world is sure to have a unique voice, but usually people will not pay for his singing. In other words, if a geometric figure is drawn to illustrate, the demand curve of his singing is also sloping downwards, but it is located in the negative value field of the vertical axis of price. In fact, the magnitude of monopolistic rent of talent does not necessarily depend on whether the demand curve is inelastic, but also on the height of it.

Attention: the market value can be subjective and not necessarily in line with the moral or objective standards. For example, some entertainment stars are rather controversial (such as the Kardashians). Quite a number of people scoff at this kind of talent, but they do be popular in the market.

Anyone is more or less unique in some aspects. As long as he finds an appropriate way to meet the needs of the market, there will be more or less monopolistic rent. As mentioned in the previous lecture, some seem to have higher profit margins only because of his unique vision or tactics, and the higher income is the return on his vision or tactics and is his (monopolistic) rent.

Secondly, not only do some persons have talents, but also locations do. For example, why must we pay for visiting a scenic spot? Because the natural scenery is especially beautiful and magnificent, or famous historical events had taken place there, or famous historical figures had lived there, there is monopolistic rent for that particular spot. Another example is mineral deposit where the mineral reserves are especially rich, or the quality is especially good, or the mining is especially easy.

Section 3: The Cause of Monopoly (II): Acquired Construction

(2) The cause of monopoly is acquired construction by human.

For example, a producer has mastered a technology that the others do not know which can reduce the direct cost or improve the quality of the product, so he can seize the market share from competitors with lower price, or keep the demand of consumers even increasing the price and obtain higher income. Obviously, the monopolistic rent here is a return on the technology.

However, competition is everywhere. If one producer uses technical barrier to exclude competitors from the market and enjoys monopolistic rent, the other producers will compete to acquire the technology or invent better technology to build their own technical barriers.

There are several ways to protect technology. One common way is that although the technology is not a high-tech one, mastering and using it needs huge funds. In reality, with transaction cost, not everyone can easily borrow enough funds to acquire the technology, which is enough to block a group of competitors. Another way is that the technology involves complicated chemical change. As long as it is kept as trade secret, it can be protected for a long time. The third way is that the technology involves physical change. In principle, competitors can use reverse engineering to acquire the technology. This kind of technology has to be protected by the patent law. Broadly speaking, although trademarks and copyrights are not the technology with physical change, they can also bring more or less monopolistic rent to producers, and are also extremely easy to be copied. Thus, a broader intellectual property law is applicable to protect them.

However, government establish entry threshold for producers not only to protect their technologies with intellectual property laws, but also to protect their privileges with administrative license, which means there is entry control in the market, and this kind of monopoly is called "administrative monopoly".

Administrative monopoly is fundamentally different from the monopoly stemmed from technology. Technology can benefit the society by reducing cost or improving quality, and should be rewarded in the form of monopolistic rent. If the laws do not protect the monopoly from technology, it is similar to infringe the PPRs of ordinary goods, and no one is willing to invest in R&D which can create new social wealth. By contrast, what administrative monopoly protects is not the activities to create new wealth, but just set up a barrier to keep the outsiders from entering the market to compete, thus creating monopolistic rent out of thin air.

If the entry threshold is constructed by technology, in order to protect their monopolistic rent, the producers already in the market will continue to develop new technologies and establish brand image, all of which can create more wealth for the society. On the contrary, administrative monopoly comes from government's license of permission for entry, which will only result in producers trying their best to favor government, more precisely, government officials who have the right to issue licenses. How to favor? Naturally, producers are willing to give some of the monopolistic rent to the officials as bribes in exchange for the license, which is precisely the cause of corruption and bribery. In economics, it is called rent-seeking, because the essence of corruption and bribery is the phenomenon of seeking the rent generated by administrative monopoly.

(3) The cause of monopoly is information cost.

In this case, one producer can search price because he has more sufficient information than the consumers, and the higher income is actually the return on the information he has, which is the rent of information.

In the market structure of price-searching caused by information cost, the number of producers is often very large, not just one, which proves that defining the concept of monopoly by the number of producers is a serious mistake. This kind of cause of monopoly will be explained in the next lecture about information cost.

Section 4: The Ambiguous Antitrust Law

After understanding the causes of monopoly, it is easy to understand why the "Antitrust Laws" are entirely wrong, and the consequence of the implementation is inevitably disastrous.

Firstly, antitrust laws invariably use the so-called "market power" as the judgment of monopoly, which means whether the number of producers in the market is large, and whether one of them has a large market share, thus having great power to influence the market price. As has discussed in Section 1, the logic of this judgment is mistaken. The number of producers in the market does not matter. What matters is whether there is entry threshold and how high it is.

There is a fallacy of so-called "Natural Monopoly". It is believed that because there are huge fixed costs in some industries, the AC decreases along with the increase of output, and there is only one producer in the market, which is the most economical. With the always decreasing AC, the larger the output, the lower the AC is. Therefore, it is the most economical to let one producer to produce all the output of the whole industry. Monopoly thus becomes an inevitable result.

However, as has discussed in Lecture 15, the AC is determined by the price, and the AC curve is always coincident with the demand curve and has nothing to do with the fixed cost (actually it should refer to the overhead cost), so there can be that the fixed cost is so large that it keeps falling when the output increases. The correct analysis should be due to the market capacity is too small, one producer is still in the part of declining ADC, the demand has reached saturation, so the market can only feed one producer.

What is more, what matters is if there is entry threshold, not if the ADC keeps falling. As long as there is no entry threshold, or the entry threshold is low enough, the potential competitors outside the market who can lower the direct cost than that of insiders will overcome the threshold to enter the market. There is no need for government who is often ignorant to the market to enact any antitrust laws.

These redundant laws are more likely to be used by the losers in the competition to attack the winners. Do not forget that human is self-interested, and those who lose in the competition will seek the government's intervention to help them if possible.

Secondly, even if it is a monopoly, as long as it is not administrative monopoly, it should not be struck. The monopoly caused by innate talent such as sport stars or scenic spot is obviously good thing. As for the monopoly caused by technology, why should it be struck? If one cannot be rewarded by the monopolistic rent, he would not bother to innovate or establish famous brand. Without new technology, how can a society make progress?

What is more, those who can only make up for the direct cost and have no rent or overhead cost are marginal producers who will be eliminated at any time due to a drop in income. A society that is made up entirely of such marginal producers and is full of the so-called pure competition market will be very fragile, unable to withstand any shock, prone to bankruptcy of large number of firms, unemployment of workers and plummeting production.

In fact, "Core Competence", which is highly valued in management science, is precisely another name of monopoly, because the key feature of it is the difficulty to be imitated or copied by competitors. In other words, what management science tries to teach producers is how to gain monopoly in the market competition. If monopoly should be struck, management science should be banned because it "instigates" producers to commit crimes which are against the antitrust laws.

For the monopoly of which the threshold is constructed by technology, it is actually formed through competition, so it cannot eliminate competition. Even if a monopolist in the market has built up a too high barrier that no outsiders can overcome it to enter the market, the rivals can have another competition way of creating a new market with subversive technology to eliminate the old market which has been highly monopolized.

The bankruptcy of Kodak is precisely a typical example. This giant had once accounted for more than 90% of the film market and 85% of the camera market in the United States, ranked among the top 500 in the world, had numerous patented technologies, and was undoubtedly a monopolist. However, struck seriously by the popularity of digital camera, the price of Kodak's shares had fallen to be delisted from the New York Stock Exchange, which means the share price is lower than US\$ 1, and at last had to apply for bankruptcy protection.

In conclusion, producers are always struggling at the edge of life and death as long as it is not administrative monopolist, because competition is everywhere, and cannot be eliminated by monopoly. There is competition in the market, for the entry threshold of the market, and even in the form of creating another new market to replace the old market highly monopolized. If one thinks he has successfully monopolized the market and can rest on the current brilliant achievements, he is only digging his own grave. One day when he wakes up, he will suddenly find the world is beyond recognition, and his monopolistic rent has been eroded or even wiped out by new technologies or new markets. There is a Chinese saying, "when rowing upstream, if you do not advance, you will drop back", which is especially true in market competition.

Section 5: Administrative Monopoly

Administrative monopoly is quite another matter. Since the entry threshold of administrative monopoly is not built up by technology or brand, but by license issued by government, under such constraint, a monopolist in the market naturally has no motivation to innovate or establish brand, but only care about engaging in public relations to favor government, more strictly speaking, government officials.

Here, the "Internet Phone Case in Fuzhou" is taken as an example to illustrate that an administrative monopolist is not only lack of motivation to develop new technologies, but also lack of interest in adopting new technologies developed by others.

In March 1997, Chen, a resident in Fuzhou downloaded a software of internet telephone (IP) and connected his computer with the residential telephone through a modem to set up an IP. In September, he used the IP as a promotion for the home appliance store run by his younger brother. Anyone who had bought home appliances in the store could use the IP to talk to their relatives or friends abroad for five minutes free of charge. Some customers asked for longer call time, so his brother applied for a public telephone in October, set up it as an IP, and began to provide long-distance telephone service. On December 23, Fuzhou Telecommunications Bureau reported a case to the police station that Chen brothers was illegal to provide international long-distance telephone service by violating the law that long-distance communication service and international communication service should be only licensed by the relevant administration of central government. The police station then distained the computers and accessories used by the Chen brothers for IP, the business income of RMB50,000, and once restricted their personal freedom for the crime of illegal business operations in the criminal law. The Chen brothers later entrusted a lawyer to bring an administrative lawsuit to the court against the police station, but was rejected in the first instance judgment. The Chen brothers appealed again to the intermediate court and finally got the result of rejecting the first instance judgment on January 20, 1998. It is the sensational "IP Case in Fuzhou" which was regarded as "the No.1 Case of Internet in China" of that time.

However, on the next day after the intermediate court in Fuzhou ruled that "IP is not an exclusive telecommunication service", the officials from the Telecommunications Administration Bureau of the Ministry of Information Industry urgently declared that "IP does be an exclusive telecommunication service", and is just "not provided for the time being".

Private IP business is illegal, because according to the authorization of the State Council of China, only China Telecom and China Unicom could engage in basic telecommunication business, and only China Telecom could engaged in international communication business. At that time, there had been many similar cases in Shanghai and Guangdong, and those involving large amounts of money were sentenced as the crime of illegal business operations in the criminal law. As China does not have a case law system, except for the judicial interpretation of the Supreme Court, the judgment of any local court has no absolute authority, so the judgment of the intermediate court in Fuzhou could not change the situation that China Telecom had monopolized the IP service but did not provide it.

From the above brief introduction of the "IP Case in Fuzhou", it is easy to find that the Chen brothers were not particularly professional technicians, but even they had no technical difficulty in providing IP business. It means that there was almost no technical barrier for this market. Then how could be the outsiders prevented from entering the market if government had not granted monopoly power to telecom companies?

What is more, how can a telecom company have not the technology that even a civilian personnel like Chen brothers can have? There is a meaningful sentence in the written verdict of the intermediate court in Fuzhou, "IP service will not harm anyone, nor is it a threat to national security, but it is a threat to traditional long-distance telephone service, which is the threat of advanced technology to out-dated technology." In fact, IP was invented as early as 1995. Why did telecom companies not adopt it? The following data can well explain why: the price of international long-distance call from China to the United States was RMB18.4 per minute at that time, while the highest price of IP by the Chen brothers was just RMB9! The difference between them implies how huge the administrative monopolistic rent of the telecom companies is.

Because the "IP Case in Fuzhou" aroused widespread concern, the telecom companies were under great pressure from the public and finally provided IP service in 1999. It also prompted the Ministry of Information Industry and the supreme court of China to issue the judicial interpretation approving the intermediate court in Fuzhou. It has been the prelude to the dramatic price drop in long-distance telephone, and a milestone event marking a turning point.

In conclusion, an administrative monopolist is unwilling to adopt new technologies with lower costs because he wants to continue to earn high monopolistic rent. Some developed countries conduct cost accounting for public utilities with so-called natural monopoly in order to prevent over-pricing, which will only make the administrative monopolists even less motivated to develop and adopt new technologies to reduce cost that will directly cause price to be reduced.

Therefore, there is no more effective way to intensify competition than abolishing license of entry control which can force the administrative monopolists to try to reduce cost and improve quality. It is administrative monopoly that most needs the antitrust laws to suppress, which has never happened. The antitrust laws are enacted by government, while the administrative monopoly is licensed by government, so how can government suppress itself?

The rent of administrative monopoly comes from entry control of government, not from technological innovation and brand building of producers, which will only make the monopolists lose competitiveness. The competitiveness of a nation is combination of that of all producers in it, so administrative monopoly will eventually damage the nation's competitiveness. There is no administrative monopoly in the international market that can be licensed by a nation. Competition is still everywhere. A nation can eliminate competition in the domestic market by administrative monopoly, but it cannot do so in the international market. In the end, a nation full of administrative monopoly will become weak in international competition, then how can the producers of it survive in the international market?

Section 6: The Deadweight Loss of Monopoly

As mentioned in the previous sections, except for administrative monopoly, normally monopoly (price-searching) is actually good for the society, because the cause of it is entry threshold which is established by innovation of new technologies, and cannot eliminate competition.

In MES, monopoly is believed to be evil based on some seemingly plausible analysis which think monopoly will cause so-called "deadweight loss", an efficiency loss shown in Fig.16-1. The demand curve in it is sloping downwards, reflecting that the market structure is price-searching (monopoly), and a monopolist can increase the price along the demand curve by decreasing the output.

If a monopolist sets a single price for all products sold, his average revenue (AR) curve will coincide with the demand curve, as has been explained in Lecture 15. His marginal revenue (MR) curve will no longer coincide with the demand curve, but is located below it, because if he increases output which is the increase in market supply, there will be a decrease in price (not only the price of the increased output, but also that of all output). Thus the increase in revenue (i.e. MR) by producing one more unit output is lower than the price (i.e. AR).

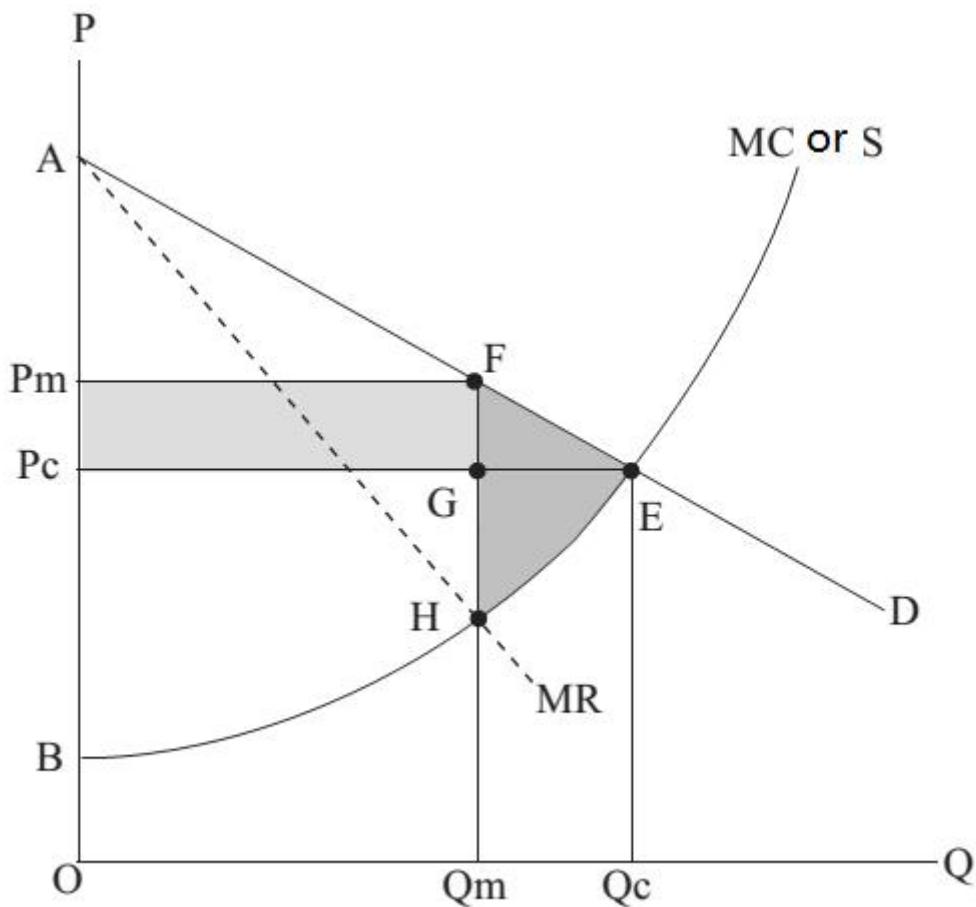


Figure 16-1

The marginal cost (MC) curve is sloping upwards in Fig.16-1, which is also the supply curve of the monopolist. Suppose there is only one producer in the market, so the supply curve of him is also that of the whole market.

According to what has been introduced in Lecture 15, the optimum for any producer in any market (no matter the market structure is price-taking or price-searching) is $MR = MC$, which is the point H in Fig.16-1. It determines the optimal output is Q_m . With the output (supply) of Q_m , the optimal price determined by the demand curve is P_m that is corresponding to the point F vertically above H.

In MSE, monopoly is believed to be evil because the optimum of it (Q_m, P_m) is compared with that of price-taking (Q_c, P_c) that is the intersection point (E) of the demand curve and the supply curve (coincident with the MC curve). The equilibrium point of monopoly or price-searching is lower in output and higher in price than that of price-taking, which is unfavorable to consumers.

However, the analysis of MSE is much more comprehensive with the consumer surplus and producer surplus as analysis tool introduced in Lecture 15. Monopoly is unfavorable for consumers but favorable for producers, while the whole society is made up of both consumers and producers, so it is not possible to draw an intuitive conclusion on whether monopoly is favorable or not for the whole society. Only by considering the changes in the benefits of both parties can an objective conclusion be drawn. Consumer surplus can be used to measure the benefit or welfare for consumers in a transaction, while producer surplus for producers. By comparing the magnitude of them, we can analyze the welfare effect of monopoly on the whole of society.

The marginal cost (MC) curve is sloping upwards in Fig.16-1, which is also the supply curve of the monopolist. Suppose there is only one producer in the market, so the supply curve of him is also that of the whole market.

According to what has been introduced in Lecture 15, the optimum for any producer in any market (no matter the market structure is price-taking or price-searching) is $MR = MC$, which is the point H in Fig.16-1. It determines the optimal output is Q_m . With the output (supply) of Q_m , the optimal price determined by the demand curve is P_m that is corresponding to the point F vertically above H.

In MSE, monopoly is believed to be evil because the optimum of it (Q_m, P_m) is compared with that of price-taking (Q_c, P_c) that is the intersection point (E) of the demand curve and the supply curve (coincident with the MC curve). The equilibrium point of monopoly or price-searching is lower in output and higher in price than that of price-taking, which is unfavorable to consumers.

However, the analysis of MSE is much more comprehensive with the consumer surplus and producer surplus as analysis tool introduced in Lecture 15. Monopoly is unfavorable for consumers but favorable for producers, while the whole society is made up of both consumers and producers, so it is not possible to draw an intuitive conclusion on whether monopoly is favorable or not for the whole society. Only by considering the changes in the benefits of both parties can an objective conclusion be drawn. Consumer surplus can be used to measure the benefit or welfare for consumers in a transaction, while producer surplus for producers. By comparing the magnitude of them, we can analyze the welfare effect of monopoly on the whole of society.

Under the market structure of price-taking, the equilibrium price is P_c , and the consumer surplus is AP_cE . By contrast, under the market structure of monopoly or price-searching, the equilibrium price is P_m , and the consumer surplus is AP_mF . Obviously, the consumer surplus has reduced by the trapezoidal area of P_mFP_cE .

Under the market structure of price-taking, the equilibrium price is P_c , and the producer surplus is P_cEB . By contrast, under the market structure of monopoly or price-searching, the equilibrium price is P_m , and the producer surplus is P_mFHB , which has one more rectangular of P_mFGP_c but one less curved triangle of GEH . Therefore, whether the producer surplus increases or decreases depends on whether the area of the rectangle larger or smaller than the curved triangle. There is no geometric theory to ensure that the rectangle must be larger than the triangle, which proves mathematically that the only producer in the market is not necessarily a monopolist. If the rectangle is smaller than the triangle, the only producer in the market would better take the price of P_c instead of the monopolistic price of P_m . However, in MSE, this possibility is completely ignored because it implies the definition of monopoly is wrong, and it is supposed that the rectangle must be larger than the triangle to ensure that producers will only choose the monopolistic price.

Based on the above analysis of consumer surplus and producer surplus, it is clear that the rectangular of P_mFHB comes from part of the trapezoidal P_mFP_cE . In other words, that part of the decrease in consumer surplus does not lose, but is transferred to producer surplus, because the rise in price increases the producers' revenue. Although this transfer is unfavorable to consumers, it is favorable to producers. From the perspective of the whole society, it is only income redistribution and does not cause any loss.

However, there is still a triangle of FGE that is not transferred to producer surplus. What is more, although the increase in producer surplus is transferred from consumers, the decrease in producer surplus (the curved triangle of GEH) is not transferred to anyone else. Thus, the curved triangle of FEH summed up by these two triangles is called the "Deadweight Loss", which is precisely the reason for MSE to criticize monopoly, because from the perspective of the whole society, it is a net loss.

The same conclusion can be made from another perspective. Any equilibrium in economics must be in line with the equal marginal principle.

For example, in the equilibrium of consumer, the marginal substitution rate of different goods is equal to their relative price that is actually the marginal cost of goods for consumer. In other words, the equilibrium should be the intersection point of demand curve (reflecting the marginal use value or MV of the good for consumer) and the price (reflecting the marginal cost of the good for consumer). Another example is that in the equilibrium of producer, marginal revenue is equal to marginal cost ($MR=MC$).

The above two examples are respectively from the perspective of consumer or producer. From the perspective of the whole society, the equilibrium of consumer and that of producer should be the same, which means in the equilibrium of optimum of the whole society, the MV of consumer is equal to the MC of producer. The economic implication is that for the whole society, when a good is produced and then consumed, the value (revenue) is consumption enjoyment (use value for consumer), and the cost is production cost, and they must be equal on the margin in equilibrium or optimum.

Under the market structure of price-taking, the intersection point of the demand curve and the supply curve is the equilibrium, and the MV represented by the demand curve is naturally equal to the MC represented by the supply curve, so the society is in optimum. However, under the market structure of monopoly or price-searching, in equilibrium, MV is at the point F, while MC is at the point H, so the former is located higher than the latter, and the society is not in optimum. The reason is that the consumption or the output is too low. By increasing consumption or output, according to the law of diminishing marginal use value, MV will decrease; while according to the law of increasing marginal cost, MC will increase, which will make them be equal in the end (the point E in Fig.16-1), and the society is in optimum.

It is the critique on monopoly in MSE: the consumer optimum and the producer optimum are not the same equilibrium, which causes the society to be not in optimum, resulting in the deadweight loss of the curved triangle (FEH).

Section 7: Turning Consumer Surplus to Producer Surplus

It seems the critique on monopoly is reasonable. However, if think more deeply, there is a great logical contradiction in the critique: how can optimum be not equilibrium? As has been explained in Lecture 7, equilibrium and optimum is actually the same, just from different perspective. If it is said that the consumer optimum and the producer optimum is not the same equilibrium, but under the same constraints, how can it happen? Different equilibriums can only be due to different constraints. Under the same constraints, both consumer and producer are in optimum, then how can the society be not in optimum?

This logic paradox implies there is serious mistake in the critique on monopoly in MSE. As has been pointed out in Lecture 7, equilibrium, as a thinking tool, can help to judge whether a certain theory or explanation is right. With the help of equilibrium or optimum, we can notice there must be mistake in the critique on monopoly in MSE, but what exactly is the mistake? The answer is: the equilibrium believed by MSE is in fact not equilibrium at all! More precisely, the situation of cost and income, supply and demand in reality is not as simple as that described in Fig.16-1.

From the perspective of consumer surplus and producer surplus, since there is benefit that no one gets (the so-called deadweight loss), why do self-interested human not try to get it back? In other words, there is deadweight loss, which implies a contradiction with the postulate of self-interest. It is logically similar to say there is a hundred-dollar bill on the ground in public, but no one tries to pick it up. Obviously, if the monopolist increases output to Q_c , and decreases the price of the additional output (from Q_m to Q_c), so as to attract consumers to buy, the deadweight loss of FEH can be recovered (the part of FGE goes to consumers, while the part of GEH goes to the monopolist).

The mainstream economists will argue, "The monopolist will not do this, because if he increases output, the decrease in price will affect all outputs, not just the additional one. Then although he can gain more of GEH, the decrease in price will cause him suffer a loss of P_mFGP_c , the former is less than the latter, which is more a harm than a good to the monopolist."

A convincing rebuttal to it is: Why must the monopolist only charge a single price for all products, rather than try to keep the same price for the original output, and only decrease the price for the additional output, so as to get the triangle of GEH back?

The mainstream economists will argue, "The monopolist will not do this, because if he increases output, the decrease in price will affect all outputs, not just the additional one. Then although he can gain more of GEH, the decrease in price will cause him suffer a loss of $PmFGP_c$, the former is less than the latter, which is more a harm than a good to the monopolist."

A convincing rebuttal to it is: Why must the monopolist only charge a single price for all products, rather than try to keep the same price for the original output, and only decrease the price for the additional output, so as to get the triangle of GEH back?

Anyone goes out to observe the market in reality can find that many producers will not only charge a single price for all output as MSE thinks. In other words, in reality, it is always in equilibrium or optimum, which is not what MSE criticizes.

As long as the preconceived idea that only a single price is charged for all the output is abandoned, there are actually many ways to eliminate the so-called deadweight loss of FEH, and the monopolist can even obtain more revenue than the loss of producer surplus (GEH). As a matter of fact, the monopolist is the one who will try his best to eliminate the deadweight loss, because he can make price, which means he is more convenient (with the least transaction cost) to do so than consumers.

Pricing along the demand curve

The monopolist can charge a different price for each unit of product. More precisely, he can charge the highest price that a consumer is willing to pay for that unit of product. The MR curve and the demand curve coincide completely, so in the equilibrium of $MR = MC$, the optimal output is exactly the same as Q_c under the market structure of price-taking. Except that all the benefits of the exchange (the area of the curved triangle of AEB in Fig.16-1) go to the monopolist, the situation is the same as the equilibrium of price-taking, so there will be no deadweight loss.

Since all the benefits of the exchange become producer surplus, and consumer surplus is zero, this kind of price arrangement can be also called "Completely Turning Consumer Surplus to Producer Surplus" (CTCS2PS). In MSE, it is mistakenly called "Perfect Price Discrimination". The characteristic of price discrimination is to charge different price for different people, not different price for different quantity of products.

Will consumers accept this price arrangement of CTCS2PS? The answer is yes. Although the price for each unit of product is different, it is the highest price he is willing to pay, and of course he will accept, which is a tautology impossible to be wrong.

Some may think that CTCS2PS is very unfavorable for consumers because they lose all consumer surpluses, and it is unfair. As mentioned in Lecture 5, economics is not ethics and morality, so the objective and scientific judgment standard is not fairness but efficiency. If there is such a waste of deadweight loss, it is inefficient. Although CTCS2PS is a very different price arrangement from price-taking, there is no deadweight loss, which is the same efficient as price-taking. What is more, consumers are willing to accept it. The monopolist can make price, but it is consumers who have the greatest power to decide whether or not to pay the price, which determines whether producers can survive or not.

However, in reality, CTCS2PS is rarely used because it is too troublesome (with high transaction cost). Just imagine each unit of goods in a store must be auctioned one by one to determine the highest price consumers are willing to pay, how time-consuming it is!

A monopolist can get more revenue through CTCS2PS, but this price arrangement will cause an increase in transaction cost. The more completely, the higher the transaction cost is. If the use value or rent of a good is rather low, and an auction specially organized for it causes even higher transaction cost, it is obviously not worth it. That is why in reality, auction is only applicable for few goods with extremely high value, such as antiques, works of art, land, etc. It is worthy to organize an auction that costs thousands or even tens of thousands of dollars only for those goods with value of millions or even billions of dollars.

(2) Block pricing

In reality, the price arrangement of block pricing is more common, which is also called "Incomplete Turning Consumer Surplus to Producer Surplus" (ITCS2PS). ¹ In Fig.16-2, it is a case of four-part pricing. If a consumer buys a small quantity (such as q_1), he has to pay a higher price (P_1). When the quantity purchased is more than q_1 but less than q_1+q_2 , the price for the part of q_1 is still the higher P_1 , while the price for the part more than q_1 is the lower P_2 , and so on.

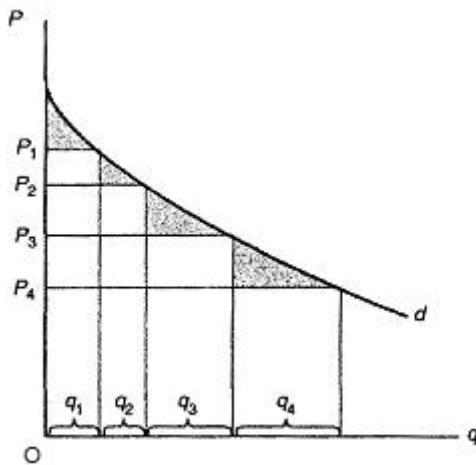


Figure 16-2

This price arrangement is actually similar to keeping the same price for the original output, and only decreasing the price for the additional output, so as to get the triangle of GEH back. As long as the output of the last part is charged the price equal to that of price-taking, there will be no deadweight loss in this block pricing, because it is only the rough version of the above pricing along the demand curve. The pricing along the demand curve is that the quantity of each part is only one unit (i.e. $q_1=q_2=q_3=\dots=q_n=1$), which means it is actually the most detailed version of block pricing. The shadow triangles in Fig.16-2 are not deadweight losses, but consumer surpluses that have not been turned to producer surpluses. That is why this price arrangement is also called ITCS2PS.

For block pricing, there are following points that should be paid attention to.

Firstly, since the monopolist only needs to charge different price for each part of goods, the increase in transaction cost is obviously lower than that of CTCS2PS where different price needs to be charged for each unit of goods. However, compared with the case of simply charging a single price for all goods, the transaction cost of ITCS2PS is still higher. From the perspective of revenue, block-pricing can turn less consumer surplus to producer surplus than CTCS2PS, but more than the case of charging a single price for all goods. Therefore, the monopolist trades off between the increasing the number of parts, thus increasing the quantity of pricing, so increasing the transaction cost, and at the same time increasing the revenue of turning more consumer surplus to producer surplus, so as to find the optimal number of parts. According to the equal marginal principle, the equilibrium is such that the increase in revenue (MR) of turning consumer surplus to producer surplus equals the increase in corresponding transaction cost (MC) by increasing the quantity of pricing.

Secondly, in reality, block pricing is easy to be confused with quantity discount. A purchase with large quantity can usually enjoy more favorable price, for example the wholesale price is generally more favorable than the retail price. It is not due to turning consumer surplus to producer surplus, but because the average direct cost is lower with larger quantity which is the internal economy mentioned in Lecture 15. For example, the price of 2-litre oil is generally higher than that of 1-litre, but on average the price of the former per liter is lower than that of the latter. It is not necessarily block pricing, but it is more likely that the bottle for 2-litre oil is larger than that for 1-litre and cost more materials, but the increase in cost is usually less than double. In other words, the average cost of packaging materials for 2-litre oil is lower than 1-litre. Under the pressure of competition, producers will let consumers to enjoy the benefits of lower cost by price reduction.

In a word, no matter whether it is CTCS2PS or ITCS2PS, as long as the price charged on the last unit of the good (the lowest price) is equal to the equilibrium price of price-taking, the monopoly's MR curve and the demand curve are still coincident, so the equilibrium point ($MR = MC$) is still the intersection point of the demand curve and the supply curve (MC curve), which means the marginal use value is equal to the marginal cost, and the society is in optimum, so there is no deadweight loss.

1. In MSE, this price arrangement is also mistakenly called as "Incomplete Price Discrimination".¹²

Section 8: The Other Price Arrangements to Eliminate Deadweight Loss

A monopolist's turning consumer producer to producer surplus is not the only price arrangement that can eliminate deadweight loss. In this section, we will introduce the others.

(3) Membership Fee or Door Money

A monopolist may find it still too troublesome to charge more than one price for different part of goods. Can he only charge a single price, but at the same time eliminate deadweight loss? The answer is yes. One kind of this price arrangement is that the monopolist charges a single price, but requires consumers to pay a membership fee before entering the door which is equal to the amount of deadweight loss (FEH). There is no deadweight loss, because it has already been recovered in the form of membership fee.

There are many clubs run in the form of membership (such as golf clubs, yacht clubs, etc.), as well as Disneyland, into which tourists must buy entrance tickets first and pay additional fees for each amusement facility. These are all the price arrangement of membership fee or door money to recover the deadweight loss.

The monopolist needs to determine the amount of membership fee or door money besides charging the single price, so there is higher transaction cost. However, as long as the revenue increase from recovering the deadweight loss by this price arrangement is higher than the additional transaction cost, it is worth it.

(4) The Pricing of All-or-Nothing

A monopolist can charge a single price, but at the same time stipulates the quantity that the consumer must buy, otherwise he is not allowed to buy. Either buy all or none, so this price arrangement is called the pricing of all-or-nothing.

The left panel and the right panel of Fig.16-3 respectively show two different pricing of all-or-nothing (P_a). The P_a in the left panel is lower than the monopolistic price (P_m), but it stipulates that consumers must buy the quantity of Q_a instead of Q_m . As long as $OA'B'$ and OAB are congruent triangles, the deadweight loss can be eliminated while the consumer surplus of $P_aBAB'P_c$ is turned to producer surplus.

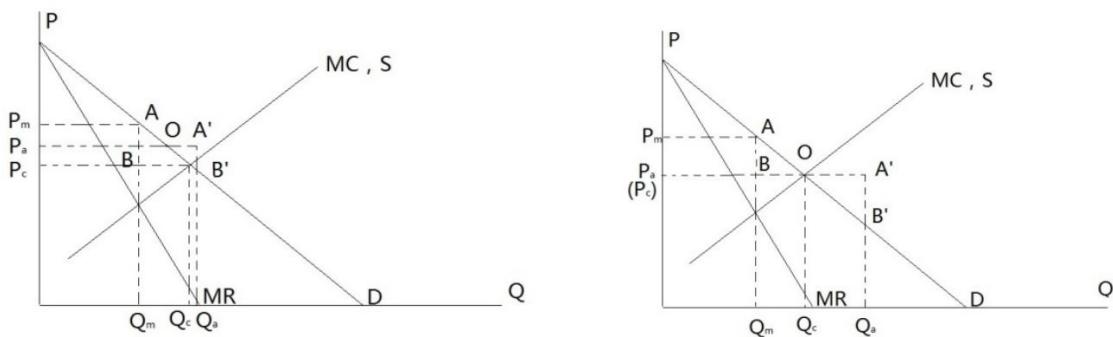


Figure 16-3

The P_a in the left panel is lower than the monopolistic price (P_m), but higher than the P_c of price-taking. The price of all-or-nothing can be equal to that of price-taking, as shown in the right panel of Fig.16-3, but the quantity Q_a that consumers must buy will be much larger than the Q_c of price-taking. As long as $OA'B'$ and OAB are still congruent triangles, the monopolist can eliminate all

deadweight loss which is the consumer surplus (AOB) turned to producer surplus.

In fact, the price of all-or-nothing can even be lower than that of price-taking! Readers can draw their own deduction by imitating Fig.16-3. Tips: OA'B' and OAB should be drawn such that they are congruent triangles, and the stipulated quantity for consumers must be very large.

As for which combination of price of all-or-nothing and stipulated quantity a monopolist will choose, it depends on whether the total revenue ($P_a \times Q_a$) has been maximized under specific constraints (including the shape and position of specific demand curve and supply curve). As shown in Fig.16-3, it is estimated visually that the total revenue of the pricing of all-or-nothing in the right panel is larger than that in the left panel.

Obviously, both the "all or nothing" pricing method and the charging of a single price and an unlimited amount are to set a price without adding any additional transaction costs. However, "all or nothing" requires a fixed amount (limited amount) in addition to pricing, so on the whole, this pricing method will still be higher than charging a single price and an unlimited amount. However, since the "all or nothing" pricing method can eliminate dead angle losses and extract consumer surplus, that is, when the monopoly earns more than a single price and is unlimited, there is no reason not to adopt it.

If a monopolist uses the pricing of all-or-nothing, he needs to determine the stipulated quantity for consumers besides charging the single price, so there is higher transaction cost. However, as long as the revenue increase from turning some of the consumer surplus to producer surplus by this price arrangement is higher than the additional transaction cost, it is worth it.

The above is a theoretical explanation for the pricing of all-or-nothing, but what is more important is there is such behavior in reality. In the market, many products are packaged. For example, the capacity of bottle for oil actually stipulates consumers to buy the specified quantity. In principle, consumers can also buy bulk oil which is almost only sold in farmer's markets without brands. Brands are precisely the entry threshold for monopolistic products, while the products without brands are in the market of price-taking where the pricing of all-or-nothing is not needed.

In addition, different packages are often available for the same brand product. Take the bottled oil as an example, there are generally small, medium and large pack sizes which are different combinations of all-or-nothing price and quantity designed by the monopolist to meet different needs of consumers. The larger pack size it is, the lower the average price per unit. The quantity discount mentioned in section 7 has a similar effect, so the influence of them in reality is often mixed.

Some may wonder: Few producers have specialized in economics. How do they know the marginal cost of a good and the marginal use value of consumers, and draw the above geometric figure to determine the combination of the all-or-nothing price and quantity?

In reality, neither do producers need to draw any geometric figure nor understand economics. What they need to do is to observe what the peers do and imitate them, or find out the suitable combination of the all-or-nothing price and quantity through trial and error.

Different brands of products with similar grades usually have the same packaging capacity. Because they are of the same grade, which means that their target customers are the same, and the demand curves for them are roughly the same in shape and position, and the most suitable combination of the all-or-nothing price and quantity for such customers should also be the same. Those who have just entered the industry can reduce the information cost by following the packaging capacity of the peers.

As for a brand-new product without peer experience, the producers can put a number of different packages on the market as trial and error. If a certain package turns too much consumer surplus to producer surplus (the area of OA'B' in Fig.16-3 is larger than OAB), consumers will be reluctant to buy because the stipulated quantity is too large. On the contrary, if a certain package turns too little consumer surplus to producer surplus (the area of OA'B is smaller than OAB), the revenue of the producer will be too low. Then the producer can gradually find out the most suitable package which is consistent with Fig.16-3.

It is logically similar to the example of idiots building gas stations mentioned in Lecture 3. The idiots do not need to be really rational, just as producers in reality do not need to really understand economics, the survival of the fittest in the market will ensure that behaviors consistent with economics can survive.

Section 9: Quality Competition

In MSE, the market structure of perfect or pure competition (price-taking) and perfect monopoly are regarded as two extremes. There are also two kinds of situations between these two extremes, one closer to perfect monopoly is oligopoly, and the other closer to perfect competition is monopolistic competition. There are limited numbers of producers in the market of oligopoly including duopoly (only two producers in the market). There are many producers in the market of monopolistic competition, but they supplied slightly different products, so they can charge slightly different prices.

However, according to the correct definition of monopoly or price-searching, the number of producers in the market does not matter, so the above distinction is meaningless. The entry threshold is the key. In a market with a higher threshold, producers (no matter one, two or a few) in it can have a wider range to search the price most favorable for them. Or else, they can only search the price in a narrower range, and can only charge a price a little higher than their competitors.

What is more, if the entry threshold is caused by information cost rather than technology or entry control by government, the number of producers in the market will be even more than that in the market of so-called perfect competition (price-taking), and there is also fierce bargaining, which will be discussed in the next lecture.

As mentioned in the previous lecture, the market structure of price-taking is very rare in reality, because it appears only when some rigor conditions are satisfied, including no difference in quality and type. In MSE, it is generally believed that a large number of products in reality differ in quality and type, thus turning the so-called perfect competition into the so-called monopolistic competition.

With different quality, producers have already faced different demand curves, so they can search price. However, products of different quality are highly substitutable to each other. Once the price difference exceeds the extent of the quality difference, there will be a large number of competitors, so the range for producers to search prices is very limited.

Similarly, with different types (such as different colors or designs), producers have also faced different demand curves, so they can search price. However, products of different type are highly substitutable to each other, so the range for producers to search prices is very limited.

For example, in China, there are all kinds of lanterns sold online. There are different materials (such as paper, cloth and yarn), and different degree of solid and beauty, so the prices are correspondingly different. Prices are different due to different quality, or strictly speaking, different costs. Generally, the bigger the lantern is, the more expensive it is. Usually the lanterns with only different colors will be not different priced, but white lanterns are a few cents cheaper than those of other colors, because in China white lanterns are only needed in funeral. In other words, there is a wider demand for colored lanterns than white ones, so there is wider range for their producers to search price.

The more popular types can be regarded as higher quality, so the difference in types can also be included in the difference in quality. Because different quality leads to different price, if government imposes price control, producers cannot compete in price, which does not mean there is no competition any longer, but price competition turns to quality competition.

For example, if the rent of a house is controlled lower than the equilibrium determined by the market, the landlord will reduce the quality of the house by poor maintenance or being not equipped with furniture and electrical appliances.

However, if the price is regulated above the equilibrium determined by the market, producers will improve the quality. A typical example is the "Sandwich War" that broke out in the aviation industry of US in the 1970s when there was quite serious regulation. Airlines which want to run a new route had to apply approval from the government department which had not approved such applications for many years. Although the government approval constituted the entry threshold, there was still competition in the aviation industry. The government even directly regulated the price of air ticket, and did not allow airlines to offer discount. However, this regulation could only restrict price competition, not quality competition. In order to get a higher attendance rate, major airlines tried their best to attract customers by providing high-quality airline meals. The government further imposed a strict rule that airline meals could only provide sandwiches, in an attempt to restrict competition, but airlines still competed to provide luxurious sandwiches.

Some may ask: isn't it good that price control leads to a rise in product quality? However, consumers do not necessarily think that the higher the product quality is, the better! Take the price control in aviation industry of US as an example. If consumers could choose, they would rather have cheaper air tickets than luxurious airline meals. In fact, when the aviation industry was deregulated in the 1980s, low-cost airlines entered the market in large numbers and offered very cheap tickets at the cost of not providing airline meals at all! These low-cost airlines quickly gained a considerable market share, which clearly shows that many consumers think that cheap air tickets without airline meals are much better than the expensive ones with high-quality airline meals.

What consumers need is not high-quality product, but the product with appropriate price and quality. It is no good for Government to impose administrative intervention forcing producers to improve quality.

Section 10: Cartels

For monopoly or price searching, there is a topic of "Cartel".

Cartel refers to the phenomenon that although there is more than one producer in the market, they collude together to make decisions like a monopolist, such as reducing output together and making the price rise from a lower oligopolistic price to a higher monopolistic price.

Of course, it again makes the mistake of using the number of producers in the market to define monopoly. However, even if this mistake is put aside, cartel is regarded as only a myth in MSE, because the members in a cartel are motivated to cheat. One member promises to abide by the agreement on output reduction, but in fact he will increase output in secret. When the price rises due to the output reduction by other members, he will benefit from both price rise and output rise. One member will do so, while other members are not idiots and will do the same, so all just promise to reduce output and secretly increase output. As a result, the output in the market does fall at all, but rise, causing the price to fall. A cartel is doomed to fail.

There is such a fable. In a village there would be a wine party. Every villager was required to bring a bottle of wine and pour it into a large barrel. One was so tricky that he thought, "Why don't I take a bottle of water? No one can know I mix a bottle of water in a large barrel of wine. I can take advantage of others by drinking their wines for paying just a bottle of water." However, when he drank the wine from the barrel, he found all was just water! Obviously, everyone was as tricky as him, so all poured water instead of wine into the barrel. The logic of this fable is the same as the cause of fragility of cartel.

In reality, there is no cartel that can survive long. The OPEC (Organization of Petroleum Exporting Countries) in the following reading material seems to be the most successful cartel. However, according to the facts provided by reading material, the reason why the OPEC seemed to be successful to raise the price of petroleum by reducing output was due to political factor of fighting with Israel, the common enemy of these member countries (mainly of Middle East). They temporarily put aside the economic interest for political interest and acted in unity. However, which really caused such ideal result, the Middle East wars that cast an uncertain shadow on petroleum output and led to an increase in the price, or their strict abiding by the cartel agreement to reduce output? It is still a question.

After the disintegration of the Soviet Union, Russia, a country with tremendous petroleum resources but not a member of OPEC, has become a major supplier of petroleum in the international market, and OPEC's influence on the price has been obviously weaken. Saudi Arabia, the largest member in OPEC, had to cut the output of petroleum the most in order to support the cartel, resulting in declining of the market share and monopolistic power.

READING MATERIAL: THE OPEC [1](#) [2](#)

Before 1960, international oil companies such as Royal Dutch Shell and Standard Oil of New Jersey were often accused of acting as a cartel aimed at keeping prices high. If so they failed – as became evident later on when the Organization of Petroleum Exporting Countries (OPEC) came into existence and really raised prices! In fact, it was the attempt of the major companies to cut oil prices that led the oil-exporting nations to establish the OPEC. (At what must have been a low point for intelligent foreign economic policy, the U.S. State Department actively encouraged formation of the OPEC cartel!)

Starting about 1960, the main OPEC nations solidified their control over pricing and production, in effect expropriating the private companies whose efforts had discovered and developed their oil resources. Thereafter private oil companies operating in OPEC countries received only what amounted to handling fees for extraction and marketing services. In 1973, for example, the Saudi Arabian government took all but about 0.60 of the 2.59 price per barrel.

The problem for the OPEC was and is to limit production. In fact, most OPEC members have not been holding back production. To the extent that the cartel has been viable, it is only because a few major producers, notably Saudi Arabia and Kuwait, have limited their own production.

The petroleum trade since formation of the OPEC has gone through several dramatic phases. After Egypt and Syria attacked Israel in 1973, the Arab countries dominating OPEC halted the previous rising trend of oil exports, aiming to influence the diplomatic policies of the Western nations. Oil prices moved sharply upward. By January 1974 the price more than quadrupled to \$11.65 per barrel. During this period the cartel was extraordinarily successful. Presumably, agreement upon foreign policy issues (mainly, opposition to Israel) helped to limit the chiseling that would otherwise have tended to undercut the cartel.

For some years afterward OPEC's power gradually weakened. Although the official price rose from 11.65 per barrel at the beginning of 1974 to 13.00 five years later (an increase of about 12%), the U.S. dollar depreciated about 38% over the same period. Thus, by January 1, 1979, the real price of OPEC crude oil was down from its peak.

But then, starting in 1979, another round of price increases was triggered by the revolution that paralyzed production in Iran. The official OPEC price rose ultimately to \$34 per barrel in late 1981. But by 1985 maintaining these high prices required increasingly severe production cutbacks by the major OPEC producers – in particular, Saudi Arabia. Owing to OPEC cutbacks and rising non-OPEC production, the OPEC share of the world market, which had been 56% in 1973, fell to only 30% in 1985.

After 1985 Saudi Arabian production began to rise, along with the exports of other OPEC members. In consequence, despite a moderate upward trend in non-OPEC output, the OPEC fraction of the world market has recovered to about 40% in recent years. The price, still at about 28 per barrel in 1985, has since ranged between 15 and \$25 per barrel – apart from a sharp temporary jump due to Iraq's occupation of Kuwait and the first Gulf War in 1990–1991.

In some periods the cartel has been hugely successful, thanks in large part to the Middle East wars of 1973 and 1990–1991 and the Iranian revolution of 1979. In the interludes between these historical shocks, two factors worked against the cartel. First, consuming nations began to use oil more economically and shift toward substitute fuels. Second, the high prices encouraged non-OPEC oil exporters such as Britain, Russia, and Mexico to expand their production and exports.

In reality, there is a cartel that is much more successful than OPEC -- trade union! Trade union is a typical seller cartel in that it gathers a large number of workers to collectively negotiate wages and other welfare conditions with the management.

As analyzed above, cartel is just a myth, but why can trade union be successful? The answer lies in the fact that trade union has the legal power to exercise violence that only police and army can have in a nation. Trade union will threaten to go on strike if they cannot reach an agreement with the management. Once there is a strike, trade union will cordon off the workplace, and no one can cross the cordon to go to work no matter whether he is a member of a trade union, or whether he agrees with the strike, otherwise he will be subjected to violence such as beating. If

trade union cannot exercise violence like other ordinary organizations, it is hard to imagine that they can maintain control over workers. Therefore, the essence of trade union is not cartel, but administrative monopoly!

As has been clearly analyzed, the entry threshold established by technology or brand cannot really eliminate competition, because technology or brand itself is something everyone can strive for. The real reason why a cartel cannot be successful is not that the members are motivated to cheat, but if there is no technology or brand as the entry threshold, a cartel agreement cannot establish a threshold out of thin air.

The self-interest of human determines that competition is everywhere, and the behavior of cheating is nothing more than the representation of self-interest under certain constraints. Only administrative monopoly, the threshold of which is established by government, can eliminate competition in the market. Therefore, the key to the success of trade union is not cartel, but administrative monopoly.

However, administrative monopoly can be protected by government for a long time, but it cannot eliminate the competition in the international market, so the cost is the competitiveness of the country in the world. In fact, those nations full of trade unions in all kinds of industries are all with weak international competitiveness. After 40 years of reform and opening up, China has become a nation with strong competitiveness in almost all manufacturing industry. One of the key reasons is the trade unions in China are not the "genuine" trade unions in western countries.

1. This reading material is an excerpt from "Price Theory and Its Application", P247-248. [↩](#)

2. The data on prices and production used here have been collected from several sources including International Economic Report of the President, Washington, DC: U.S. Government Printing Office, February 1974, pp. 110-111; Los Angeles Times (March 15, 1983), p. 1; M. A. Adelman, *The Genie out of the Bottle: World Oil since 1970* (M.I.T. Press, 1995), especially Figure 6.1 (p. 144), A. F. Alhajji and David Huettner, "OPEC and Other Commodity Cartels: A Comparison," *Energy Policy*, v.28 (2000), and Energy Information Administration, *Monthly Energy Review* (July 2003). [↩](#)

Section 11: The Fallacy of Game Theory

The conclusion that cartels cannot be successful in reality may make some think of the "Game Theory" that prevailed in economics after the World War II, because one of the most classical games "the Prisoners' Dilemma" is said to be able to explain the failure of cartels.

However, the explanation power of game theory is in fact so poor that it is not a useful scientific economic theory. In view of the seriously misleading of game theory, in this section, the most classic game of the prisoners' dilemma will be analyzed in detail and the fallacy of it will be pointed out.

There are two versions of the prisoners' dilemma that are illustrated in the payoff matrix of Table 16-1. The former one the original version of the game, and the latter one is the extension used to explain cartels. Only the former version will be analyzed in detail, because the same logic applies to the latter one.

Months of Imprisonment

		Don't confess	Confess
Panel (a)	Don't confess	- 1, - 1	- 36, 0
	Confess	0, - 36	- 24, - 24

Rank-ordered payoffs

		Small output	Large output
Panel (b)	Small output	3, 3	1, 4
	Large output	4, 1	2, 2

Table 16-1

The story of the prisoners' dilemma is as follows. The police have apprehended two men, accomplices in a crime, but the evidence against them is weak. Lacking a confession, the authorities will be able to impose only a minor penalty. But if either prisoner confesses, conviction on a major count is guaranteed. Isolating the prisoners from one another, the district attorney offers to let either one go free in return for his turning state's evidence - provided that the other does not also do so. (If both confess, each will receive a reduced punishment.)

In Panel (a) of Table 16-1, each prisoner has the strategy options Confess or Don't Confess. The rows represent possible strategies of one prisoner, while the columns represent those of the other prisoner. In each cell, the paired numbers are the payoffs: the first number is the payoff to the Row player, the second the payoff to the Column player.

If both choose not to confess, the payoff of each is -1 (one month in jail) in the upper-left cell. If both choose to confess, the payoff of each is -24 (two years in jail) in the lower-right cell. If the row player chooses not to confess, while the column player to confess, the payoff of the former is -36 (three years in jail) and that of the latter is 0 (go free) in the upper-right cell. The combination of options in the lower-left cell is the opposite.

How to find the equilibrium of the prisoners' dilemma?

Firstly, from the perspective of the row player, if the column player has chosen not to confess, which means he can only choose between the two left cells, of course he should choose to confess because $-1 < 0$. If the column player has chosen to confess, which means he can only choose between the two right cells, he should also choose to confess because $-36 < -24$. In other words, no matter how the column player chooses, the optimum for the row player is always "Confess".

Secondly, from the perspective of the column player, he does not need to consider the possibility that the row player chooses not to confess, which means he can only choose between the two lower cells, of course he should choose to confess because $-36 < -24$. In other words, both choose to confess and receive two years in jail of punishment, which is the lower-right cell.

However, the optimum for the two prisoners as a whole should be the upper-left cell, which means both choose not to confess, because the sum of their payoffs is -2 , which is the highest among these four cells. By contrast, the sum of the payoffs in the lower-right cell is -48 , the lowest among the four cells, not only worse than that in the upper-left cell, but also worse than the sum of the payoffs (-36) in the upper-right or lower-left cell. The "dilemma" here is that each does better confessing, regardless of what his accomplice does – even though they could gain by both refusing to confess. Everyone tries to seek the best result in line with the postulate of self-interest or rationality, but finally they get the worst result for the whole that seems to be irrational. The prisoners' dilemma is famous for explaining such a paradox of individual rationality leading to collective irrationality.

This prisoners' dilemma is believed to be widely used to explain many phenomena in reality. For example, why are cartels difficult to be successful? Another example is why Coca-Cola and Pepsi both compete to spend huge amounts of money on advertising, which is thought to be also a prisoners' dilemma. If both do not advertise, they will have the same publicity, so have the same market share, and save a lot of cost on advertisement, which is good for both. Why do they still compete to advertise? If one of them advertises and the other does not, the publicity of the one advertising will far exceed that of the other, thus taking away most of the market share.

Therefore, no one dares not to advertise.

Still another example of prisoners' dilemma is the arms race between the United States and the Soviet Union. During the cold war, both spent huge amounts of money to expand the armaments, especially on nuclear weapons. Once a war broke out, the nuclear weapons would only destroy each other, so no one dared to start a war, and the world was peaceful. However, if neither had nuclear weapons, neither would have the military power to start a war, the world would be still peaceful. In the former situation, both had to burden huge military expenditures, which is obviously much worse than the later situation where neither had nuclear weapons. Why should both choose arms race? If one of them had nuclear weapons while the other did not, the military power of the one with nuclear weapons will overwhelm that of the other who would have to submit to the rival in international and even internal affairs. Therefore, no one dared not to have no nuclear weapons.

However, there are serious mistakes in the prisoners' dilemma, which will be pointed out, and then the above phenomena believed to be explained by it will be also refuted one by one.

There is an implicit supposition about constraint in the prisoners' dilemma: the two prisoners cannot communicate with each other to collude not to confess so as to reach the collective optimum of upper-left cell (also known as "cooperative solution" in the game theory). But why? What is the constraint? Obviously, it is the transaction cost of cooperation for the two prisoners must be too high! They were separated physically, but why do they not bribe the guardian?

Thus, the real question is: why are the two prisoners unwilling to pay a high enough price to avoid falling into the worst lower-right cell where both confess? The answer is obvious: because the price is higher than the cost of the punishment (two years in jail). In other words, the payoffs of the so-called optimum of the upper-left cell are completely wrong in only calculating the benefit of cooperation (both choose not to confess) without the cost of cooperation. By converting the bribe paid to the guardian into the time spent in jail, the cost of cooperation could be taken into account. The payoffs in the upper-left cell must be worse than -24, otherwise the two prisoners would have chosen to bribe the guardian to reach this so-called optimal solution (equilibrium).

In this book, we have pointed out from the very beginning that there are full of mistakes in MSE, often due to the ignorance of the important constraint of transaction cost. Although game theory is a "new" theory that emerged after the World War II, it makes the same "old" mistake of ignoring the constraint of transaction cost. In fact, individual rationality will not lead to collective irrationality. Any conclusions that are out of line with the postulate of self-interest or rationality are sure to be wrong in ignoring a certain constraint, usually transaction cost.

Actually, in reality, any cell in the prisoners' dilemma can be optimum or equilibrium, as long as the constraints are appropriate. For example, the upper-right or lower-left cell may seem to be disequilibrium, but if one of the two prisoners is the gang leader, while the other is his subordinate, with the transaction cost of cooperation low enough, they will naturally choose to protect the leader by sacrificing the subordinate.

Some game theorists may argue that it has already changed the game rules, so the prisoners' dilemma is no longer applicable. However, when faced with different constraints, game theorists create a new game by modifying the game rules, which is precisely a typical example of ad hoc theory.

Next, we will make a more specific criticism one by one on the phenomena that are said to be explained by the prisoners' dilemma.

Firstly, the arms race between the United States and the Soviet Union. In reality, both have been negotiating to reduce the number of nuclear weapons, but the results are far from satisfactory, which implies how high the transaction cost of cooperation is. Just imagine: if there is a super power over the two countries (such as world government or God) which can greatly reduce the transaction cost of negotiations, the option that both do not have nuclear weapons will certainly be chosen. A simple fact is: all local governments in a nation are absolutely not allowed to have their own armies! One reason for it is that the national defense is a public good provided by the central government is the least cost. There is also another important reason is that when there is a super-power of the central government over the local governments, it is a consensus that all local governments should not engage in armament, and the transaction cost of cooperation basically falls to zero! However, when the central government is weak (for example, during the period of the last years of a dynasty in ancient China), it is very common for a local government to have its own army that is called "warlord". The change of phenomenon caused by the change of constraint strongly proves that it is the key that whether there is a super power over the game players to reduce the transaction cost of their cooperation.

Secondly, the explanation of cartels by the prisoners' dilemma is also wrong. As has been analyzed in the previous section, the key to the failure of cartels lies not in deception or non-cooperation, but in the fact that only a cartel agreement cannot really establish the entry threshold. What is more, the members or players are motivated to cheat or not choose to cooperate, which simply means that human is self-interested. It is a tautology to explain the phenomenon directly with this postulate of economics. By contrast, whether the cartel agreement can establish the entry threshold is an objective fact that can be observed and tested, so at least it is a refutable explanation and is therefore scientific.

It is precisely the problem of game theory. It just repeats the postulate of self-interest with some new terms and seemingly interesting stories, but the essence of it is a tautology. In addition, the above criticism on the prisoners' dilemma shows that it does not take the key constraints into account, but only imagines the constraints and constantly changes the type of games if the constraints change, which is a typical mistake of ad hoc theory.

Finally, there is the case of Coca-Cola and Pepsi which compete to advertise, which is even more serious mistake to use the prisoners' dilemma to explain. Is Coca-Cola advertises only against Pepsi? The game theorists have obviously made the serious mistake of defining monopoly by the number of producers in the market again! In fact, if Coca-Cola and Pepsi do not advertise, their market shares will definitely not be the same as they both advertise, but will plummet! Regardless of the fact that there are other beverage companies which are substitutes for Coke or Pepsi, even if there are not, there are still numerous competitors outside the market. If Coca-Cola and Pepsi do not advertise, both will become anonymous. Without the entry threshold established by brands, the competitors outside the market will flood in and quickly take away all the shares from them.

Through the above analysis on game theory, we hope readers can have a deeper understanding of scientific economics which is different from the engineering theory such as the production function in MSE, or the mathematical theorem such as the Nash equilibrium in game theory.

Lecture 17: Information Cost

As mentioned in the previous lectures, information cost is one kind of transaction cost. In fact, it is a very important one, but also a very difficult one. Therefore, although there has already been a lecture on "Transaction Cost", in this lecture, the "Information Cost" is still need to be singled out for specific introduction.

Section 1: Information Cost in Society

Strictly speaking, in Robinson's one-man world, there is also information cost. Robinson needs to know all kinds of knowledge necessary for survival, such as drilling wood to make fire, picking fruits, hunting animals, etc. which are all information costs. Therefore, information cost is not necessary to be transaction cost.

However, once in society, almost all information costs are transaction costs. There is only one case that information cost is not transaction cost in society: different people have incomplete information about something, but there is no information asymmetry among them.

"Incomplete Information" means that people do not have all the relevant information about something. The unknown information makes people unable to make better use of it to obtain more revenue at less cost. The increased cost and reduced revenue simply the information cost. What is more, in order to know more information, people have to spend more costs. For example, in order to know what the weather will be like tomorrow, scientists must first study the principles of meteorology, and then launch meteorological satellites to collect relevant information, which are all huge information costs. According to equal marginal principle, the equilibrium is that people choose to pay more information cost to know more information until the marginal revenue and the marginal cost of knowing more information are equal. It is impossible for people to have complete information, and be always in a state of incomplete information to some extent, because the revenue from the more complete information is not enough to make up for the cost. No matter it is in Robinson's one-man world or in society, there is incomplete information, so it is not necessary transaction cost.

"Information Asymmetry" means that different people have different information about the same thing. Some of them have more sufficient information than others, and are in a state with relatively complete information. There is a Chinese saying, "A buyer is not as astute as a seller", which does not mean a buyer is more stupid than a seller, but he usually knows less about the goods than the seller, so the seller can use the information he knows to deceive the buyer, and get more revenue. Obviously, there cannot be information asymmetry in Robinson's one-man world, so it is transaction cost.

In society, as long as all have the same information about something (there is no information asymmetry), even if they do not have sufficient information (there is incomplete information), all make decisions based on the same information, so all will do the same thing and get the same revenue, where information cost is not transaction cost. However, this condition is too rare, so information cost in society is basically transaction cost.

Information, or some economists such as Hayek use the term of "knowledge", is as scarce as ordinary goods, so it has a price, and can bring income to those who have it, which means information is an asset. The income of an asset is rent, so the income of information is the information rent which is the same as the rent of a house for the landlord.

Take the information of weather as an example. The ancients did not have the knowledge of meteorology, while agricultural production depended on the weather. In ancient times, if the weather changed suddenly, farmers could do nothing but watched the crops destroyed. By contrast, in modern times, human have the knowledge of meteorology, and the meteorological offices closely monitor the weather changes. The sudden changes of the weather one day later can be forecasted, and the farmers will be notified in time to rush for the harvest. It is information rent that the farmers' income is higher because they can know the weather change in advance.

Another example is one who pays attention to the weather change can know that it will rain tomorrow, so he buys umbrellas in advance and sells them on the street. He can sell more umbrellas than on sunny days, and even increase the price of. The higher income is also information rent, not because he is lucky and happened to sell umbrellas on the street in rainy days, so it is not unexpected profit. Compared with those who do not pay attention to the weather forecast and do not take umbrellas, he makes money by information asymmetry.

However, information rent is often not as easily to get as the rent of a house. Once an owner of the information tells it to another one, the latter will pretend that he has already known the information and refuse to pay for it. If the owner of information asks the buyer to pay first, it is difficult for the buyer to judge whether the information is worth the price. Information is more difficult to trade than ordinary goods, because the PPR of information is not easy to protect. Once one has known the information, he cannot become unaware. By contrast, an ordinary good can be forcibly returned to the owner if it is illegally taken away. The cost for protection of PPRs is transaction cost in broad sense, so the difficulty in protecting PPR of information means that the transaction cost for protecting PPR of information is higher.

In fact, information is a kind of public good. Public good is characterized by non-exclusive use which means it is difficult to exclude the use of non-owners. Therefore, non-owners can easily infringe the PPR of public good without paying to the owner, which make it difficult to directly trade public goods.

In reality, the way of protection for information is to try to concretize the information into a tangible good, and then the information can be indirectly protected by direct protection for the tangible good. Tangible goods are "private goods" that are easy to prevent infringement by non-owners, which means the transaction cost of protection for PPRs is lower. Therefore, one of the major ways to protect the PPRs of public goods is to tie sale of the public goods and private goods. In fact, the patent laws and intellectual property laws protect the tangible goods concretized from information (technology, knowledge).

In fact, there are also a lot of ways to protect the PPR of information besides laws. For example, the umbrella seller obtains the rent of information about tomorrow's rain by selling such a tangible good as umbrella. Obviously, only in the case of information asymmetry, one needs to find ways to protect the PPR of information, because those with more complete information need to sell information to those with less information.

There are a lot of people in society who live by earning information rent. The intermediaries are the most typical. They match buyers with sellers by specializing in collecting the information of buyers and sellers. One with work experience is more popular with employers than fresh graduates, and is easier to find a job with better pay. It seems because that the rent of his labor is higher, but actually because he has more sufficient information about the job, which means that his higher rent of labor than that of fresh graduates is actually the information rent. One more example is that the income of teachers is almost entirely information rent. Teachers teach students knowledge, and students pay tuition fees in return for teachers.

Although it is difficult to protect the PPRs of public good, there are still many ways. One way is to tie sale it with a private good which has been pointed out above. The second way is that even if information like knowledge is told directly to the learners, those with limited learning ability cannot master it, so learning ability becomes a natural threshold, effectively excluding a considerable number of people. The third way is indirect charge. For example, the price of TV programs is indirectly charged by advertisements: the audiences do not directly pay for watching the programs, but the advertisers do. After watching the advertisements, the audience is attracted to buy more products from the advertisers, so indirectly pay for the programs. The fourth way is that government supply the public goods (such as national defense) and makes the users to pay through its coercive force.

In MSE, public goods are believed to be only supplied by government or for free, which is obviously wrong according to the above analysis. Even if it is necessary for government to impose fees, it is not necessary for it to provide them. The government can authorize private suppliers to collect fees and provide the public goods. As the saying goes, "There's no such thing as a free lunch." Those provided by government for free are actually paid by taxpayers, and government just acts as an intermediary like the advertisers of TV programs. Unless it is a public good generally enjoyed by all taxpayers, such as national defense, those that can only be used by a specific group of people should be paid by users according to the equal marginal principle.

Section 2: The Behaviors Reducing Information Cost

Since information cost is a kind of cost, human is motivated to reduce it due to the nature of self-interest.

As mentioned in Lecture 8, the strategy of “judging quality by price” implies that the information of price is lower than that of quality, so it is a typical behavior reducing information cost. In fact, Of course, the information cost involved in the strategy of judging quality by price is not just that of price, but also including to the risk of making mistakes because risk is also a kind of information cost, which will explained later in this lecture.

The purpose of advertising is also to reduce information cost. There are two functions of advertising. One is to tell consumers directly about the product's information: what is the performance, where is it sold, what is the price, etc. The other is to establish brand. Why should a producer establish brand? As has been discussed in Lecture 16, a producer establish brand to establish entry threshold. Strictly speaking, the entry threshold is not formed by brand, but by higher quality. Brand can make consumers easily recognize that it is a high-quality product, which means it can reduce information cost.

The Soviet Union once forbade producers to use trademarks, resulting in a sharp drop in product quality. Without a trademark (the visualization of brand), consumers cannot distinguish similar products of different producers, because they are not professionals of various products and cannot judge the quality from the appearance. Thus, producers tend to reduce the quality that cannot be judged from the appearance to reduce production cost without.

Attention: many behaviors of businesses in reality seem to be quite different from advertising (such as the advertisements on mass media like newspapers, radio, television, internet), but the essence of them are actually advertising.

For example, the fancies of iPhone will rush to queue up to buy the new products, which will even attract scalpers. As pointed out in Lecture 9, price control imposed by government will lead to queuing because the price criterion are taken place by the non-price criterion of fire-come-first-served. It is a waste (rent dissipation) for buyers waiting in line doing nothing. However, this analysis cannot be applied mechanically and rigidly to the queuing of buying iPhone. There is queuing up, which implies that the price of iPhone must be lower than equilibrium. But no government imposes price control on iPhone, and the Apple Inc. is voluntary to charge a lower price.

Why does the Apple Inc. not raise the price to eliminate the queuing and turn the rent dissipation caused by queuing to its revenue? In fact, it is a behavior of advertising. The less revenue because of lower price is not a waste, but advertising expense, the effect of which is much better than the direct advertisement on mass media. The craziness of fancies of iPhone and the scalpers are widely reported as news instead of advertisement, and people believe news but may not believe advertisement. They discuss how popular the new iPhone is on major online portals and forums such as Weibo, unknowingly help to propaganda the new iPhone. The wide coverage and strong penetration are far better than the traditional advertisement, because the audiences are unconsciously attracted to actively participate in the event of marketing. This strategy of Apple Inc. is concluded as “hungry marketing” by deliberately charging the price lower than equilibrium.

There is also another example. When a new product is launching on the market, the price is often charged lower, sometimes so low that it is even lower than the direct cost. In MSE, it is believed to be "Predatory Dumping", a kind of monopolization by eliminating the original producers in the market, and raising the price to higher monopolistic price to cover the previous loss when the new producer has become the only one in the market or a monopolist. However, as has been criticized in Lecture 16, whether the market structure is monopoly depends not on the number of producers in the market, but on whether there is entry threshold. If a new producer drives out the original producers by charging a price lower than direct cost without establishing any entry threshold, he is foolish to think that the original producers who have left will not enter the market again when he raise the price to higher monopolistic price. Therefore, the so-called predatory dumping is again a myth.

Then why does a new producer charge the price lower than the direct cost? Because it is again a behavior of advertising. A new product is not well known to consumers, so it needs to be sold at an ultra-low price (sometimes even for free) to attract consumers' attention and encourage them to try it to experience the function and performance of it, which is much better than direct advertisement on mass media. Consumers who will buy it or get it for free are usually the target customers of the product. The loss caused by charging a price lower than the direct cost is equivalent to the advertising expense, but it is a much more accurate marketing than traditional advertisement.

In conclusion, there are a large number of cases about artful advertisement in marketing and advertising. There seems to be waste or loss which in fact is advertising expense and is better to reduce information cost of making consumers know the products or the brands.

Thus, the purpose of advertising is to reduce the information cost of consumers, and it is the product quality that can really establish the entry threshold. Although there is always false publicity, the true information will spread out, and the information cost will gradually decrease at last. If a product without good quality is only advertised by cheating consumers, which means the advertisement does not reduce but increase the information cost of consumers, or the entry threshold is established by information cost instead of product quality, the information rent will be dissipated by the market competition sooner or later.

The case of "Qin-chi", a brand of Chinese liquor, is a typical negative example. Qin-chi became the bid winner of CCTV in 1996 and 1997. In 1996, the winning bid price is 66.66 million, and in 1997, it soared to the sky-high price of 320 million. Initially, it had won a great reputation and outstanding achievements within one year, and became famous in the whole country of China. In 1996, the sales volume increased by more than 500%, and the profit margin increased by more than 600%, but the production capacity finally reached the limit, and could not keep up with the rapid expansion in 1997. Then it purchased other liquor in Sichuan Province to blend with its liquor, which is in fact at the cost of reducing the quality. The public had doubted how Qin-chi could bear such a huge amount of advertising expense, so news media soon dig into the fact and exposed this blending practice. Since then, the situation had taken a sharp turn for the worse. In 1997, the sales volume dropped 300 million compared with that in 1996, and the profit and tax fell 60 million. In the first four months of 1998, the sales volume further dropped 50 million. The brilliant performance turned into a bubble that burst in a flash.

Section 3: The Behaviors Increasing Information Cost

Since information asymmetry can bring information rent to those have more information, while it is information cost for those who have less information, there will be also behaviors increasing information cost. Previously, it is the behaviors of self-interested human reducing various costs including transaction cost that is discussed, but if increasing information cost can bring information rent, self-interested human will also do so.

There are two kinds of such behaviors increasing information cost.

(1) The concealment of information

Suppose the story of the movie "2012" happens in reality that all except one do not know the end of the world is coming. The One who knows can choose to inform the public so that all work together to build many Noah's Arks to save as many people as possible. And he can also choose to conceal the information and quietly build Noah's Arks for sale when the end of the world comes. What will he choose without the consideration about morality?

As for the former option, he cannot get income from the disclosure of information, while as for the latter option, he can profit from concealing information. From the perspective of the whole society, if he chooses the former, all can work together to build more Arks to save more lives, so the decrease in information cost by the disclosure of information is greater than the information rent brought to the one who knows by the concealment of information. However, he cannot get the benefit of reduction in information cost. Imagine that if he can get considerable part of the benefit from the society through negotiation, he will choose the former option. The question is how he can negotiate with the society if he does not disclose the information? So it goes back to the problem mentioned in Section 1: the transaction cost for trading information that is essentially public good is so high that the owner tends to protect his PPRs of information by concealment.

It is morality as another institution that plays a role in reducing transaction cost. The one who knows cannot get any monetary income from disclosing information, but all will respect him as a saint, which is very high non-monetary income.

On the other hand, the market as an institution has actually played a role in reducing transaction cost. If in a society, the PPRs of a good were not protected, the one who knows would even do not have the means to get information rent by selling Noah's Arks, so he would still not tell others about the information, but quietly build only a Noah's Ark to carry his loved ones to escape with him. As a result, fewer people would survive the end of the world.

Because this example involves the destruction of human beings, the psychological pressure of concealing information is very great, and it will not affect the escape of the one who knows if he discloses the information, which should be enough to overwhelm his self-interest. However, in some less extreme cases, concealment of information is often the optimum.

Milton Friedman once explained the function of PPRs and the market in reducing information cost. Suppose one knows how to grow grain on stones. If there were no PPRs and thus no market, he would not have the motivation to disclose this information.

If there were PPRs for grain but were not for information, he would quietly make use of the information to produce grain for sale. Although he did not disclose the information directly, others observed that he had no land but only stones and could supply food, which at least made the information of "grain can be grown on stones" leaked out by his sale of food. Of course, the key information on how to grow grain on stones is still unknown, but others will compete to discover the secret, which is easier to find than completely unaware of the original information that grain can also be grown on stones.

If there were PPRs not only for grain but also for information, he would choose to sell or rent the information to those who are good at growing grain and will certainly produce more food than the one who may be only good at inventing instead of farming. The increase in the output of grain reflects the increase in information rent from the protection of PPRs of information (or the reduction in rent dissipation or transaction cost).

It shows that PPRs and the market can guide the self-interested people to disclose information, thus reducing information cost and benefiting the whole society. In fact, it has been pointed out as early as in Lecture 3 that the market can guide self-interested human to benefit himself by benefiting others.

(2) Price-making

As mentioned in Lecture 16, the term "price-making" in MSE refers to monopoly, but here it is changed to refer to the other behavior of increasing information cost by spreading false information to mislead others, or known as "cheating".

The case of Qin-chi mentioned in the previous section is a typical example of price-making. From this example, it can be concluded that price-making is not sustainably effective on the one hand, and it may eventually lead to doing more harm than good to the price-maker on the other hand. When Qin-chi won the bid of CCTV for the first time in 1996, the product quality and production capacity of it were enough to cover the cost of 66.66 million, so the sales volume and profit margin of that year greatly increased. However, the bid price of 320 million in 1997 had been totally out of line with the real capability of it.

Price-making is popular in the market of artwork. Some price-making are rather successful, because the artworks are really good, but the authors are not famous at the beginning, and need to attract the attention by price-making. When the quality of the artworks has been recognized, the high price made will not fall. In other words, the successful price-making is actually a behavior reducing information cost. However, most of the price-making in artworks is not successful in the long run, which is equivalent to the case of Qin-chi. The high price is out of line with the real quality or capacity, so it is unsustainable. Due to the successful price-making, many believe that only by price-making, they can make money out of thin air. After all, Qin-chi in 1997 was so crazy because it was cheated by the super success in 1996.

Therefore, the problem still lies in information cost. If information cost is low enough, the artworks or firms with good quality can be discovered immediately without publicity, and they will be immediately and accurately priced by the market, there will be no need for price-making. Or no one will be cheated, and the price-making cannot be successful even in the short run, so there will be no room for price-making.

In a word, self-interest is good in that human will try to reduce cost (including transaction cost), while it is also harmful in that human will also try to increase cost which is precisely transaction cost. However, self-interest can be harmful just because there is transaction cost.

Section 4: The Price-Searching Caused by Information Cost

As mentioned in Lecture 16, one of the causes of monopoly or price-searching is information cost that constitutes the entry threshold of the market. This cause of monopoly is very special, because there are too many producers in the market, which means the number of producers is even more than that in price-taking.

In tourist attractions, there are many stalls selling local specialties that are almost the same, so it is hard to say the producers are monopolists. However, as is known to all (including even the tourists), the prices of the goods there must be higher than those in local ordinary market. Most tourists are still willing to pay higher price because they have to hurry to the next scenic spot and have no time to go to the local ordinary market. It is the rent of the stalls in tourist attractions, partly from the saving the information cost for the tourists to look for the specialties in the local ordinary market.

What is more, there is always bargaining in these stalls selling local specialties, which means both buyers and sellers are searching price! If it is a monopolistic product defined in MSE, bargaining is understandable. But there are so many producers selling similar products, and it seems to be the so-called pure competition, how can be there bargaining? Because it is a local specialty, there is high information cost for tourists as outsiders. On the one hand, tourists will judge quality by price, so if the price is too low, they will think the quality is poor, which will reduce the sales volume and the income. On the other hand, different tourists have different information, so the sellers quote a higher price at the beginning to make room for bargaining. Through bargaining, the sellers can judge how much information the buyers have, according to which, he can adjust the price to charge different price for different buyers with different information. In other words, it is price discrimination.

Why can the producers who allow bargaining and implement price discrimination win those who do not allow bargaining and charge a lower price on average? Because after bargaining, some tourists pay a lower price than that in price-taking, which grabs away the business of the producers charging a price equal to that in price-taking, so they cannot have the optimal quantity and are eliminated. It implies that their pricing is not low enough. But if the price is too low, tourists will think the good is poor in quality, so the sales volume may drop further. As for some tourists who have paid higher price than that in price-taking, they believe that they have bought the good with better quality, which may not be wrong due to the information cost.

However, because the price is quoted higher at the beginning, it is inevitable that some buyers will be scared away, and the number of buyers will be less. Thus, there will be idle salesman, which means that there are too many producers in the market. If there are fewer producers, each producer can serve more customers, and there will be no idle salesman. It seems to be a waste that there are idle salesman, but the reason is information cost, which causes producers to implement price discrimination through bargaining, thus causing there are more producers in the market than those in price-taking without information cost.

Section 5: Replacing Risk with Information Cost

Another major application of information cost is the replacement of the concept of risk.

F. H. Knight (1885-1972) distinguished between "Uncertainty" and "Risk", but now there is no distinguish between these two concepts in economics. However, there is problem in the concept of risk that it is not observable and refutable. Just like the critique about the concept of equilibrium, look out of the window, can one tell whether there is more or less risk in the world? The concept of equilibrium can be partially saved by assuming that the world is always in a state of equilibrium, but the concept of risk cannot be saved like that, because it is not important to know the world is always in a state of risk. What is important is that when applying the equal marginal principle, we must know the change in constraints. Specifically for risk, we must know whether there is more or less risk with the change in constraints, while it is no need to know there is more or less equilibrium.

It is better to replace the concept of risk with information cost. There is risk, which means people cannot predict the future 100% accurately. In other words, there is information cost which makes people unable to have complete information.

There is "Determinism" in philosophy, which holds that there are objective laws and causality in nature and society. As long as all previous conditions are fully known, the future can be predicted completely and accurately. As a determinist's saying, my hand is now on my left leg instead of right leg, which has been decided at the moment the universe was formed. Whether determinism is true or not is a philosophical issue, but in economics it can be understood that if one can know all the key constraints that determine the future, he can infer, or not very precisely speaking, predict the future, because it is logically similar to explaining the past. Thus, people will make mistakes in prediction because they do not know all the key constraints in time, which means there is information cost.

Unlike the change in risk that is not observable and refutable, information cost is at least observable at the margin. For example, in modern times, farmers have more sufficient meteorological knowledge and information than those in ancients, which means the information cost on agricultural production has decreased. Obviously, it is more observable than the statement that the risk of agricultural production has decreased.

As mentioned before, theoretically Newton's law can be applied to accurately infer which side of a coin will face upward when it is thrown up and then falls down. However, physicists can only calculate the probability because there are too many constraints that change fast and affect the coin. Logically similarly, theoretically, the stock market can be accurately inferred by applying the law of demand, but economists can only calculate the probability because there are too many constraints that change fast and affect the stock market.

In fact, in MSE risk is measured by probability, which is problematic. The probability measurement is based on the "Law of Large Numbers" in statistics which states that when the test condition is unchanged and the test is repeated many times, the frequency of a random event is close to the probability of it. Therefore, if the test condition changes, or there are not enough random samples, the result will not be accurate enough.

Section 6: Insurance

For the law of large numbers, although it is not useful to predict specific events because the applicable conditions are not satisfied, it is the foundation of the insurance industry. Since risk will mislead people make wrong decisions, they are averse to risk, just as they do not like cost, which is better understood by replacing the concept of risk with information cost. Thus people are willing to pay a price in exchange for others to take risk. This price is called "risk premium" in finance and investment, which implies that there is a market for risk and information cost can be priced. In the financial market, the behavior of purchasing risk is called "speculation" which has nothing to do with the law of large numbers. However, if the law of large numbers is applied to purchase risk, it is "insurance".

For example, suppose one has 1 million worth of property. If the property is insured, once it is stolen or destroyed, the insurance company is responsible for compensating him 1 million to ensure that his wealth remains unchanged, thus transferring the risk of property loss to the insurance company. Of course, he has to pay a price for this insurance which is the insurance premium (while the one million is the insured amount).

If an insurance company can accept insurance widely, it can accurately calculate the compensation expense according to the law of large numbers. Although it is impossible to know precisely whether a specific subject-matter insured will be lost and need compensation, according to the law of large numbers, a specific ratio of subject-matters insured is sure to be lost. Thus, the information that an insurance company needs to have is not which specific subject-matter insured will be lost, but the probability of a large number of subject-matters insured to be lost. Since there is the law of large numbers that can be used to calculate the probability, the relevant information cost drops significantly. An insurance company actually does not change the risk, but change the type of information it needs to have: from the type with higher information cost to that with lower information cost.

Suppose that there are 100 persons with the same constraints, and each insures 1 million of property. The insurance company estimates that the accident rate of them is 1% according to the law of large numbers based on past experience data, which means the compensation expenditure can be calculated as 1 million. The insurance company can make up for this compensation expenditure by collecting 10,000 of insurance premiums from each policyholder.

Therefore, the essence of insurance is that all gather together to help the 1% unfortunate persons who have suffered from accidents. It does not depend on one's mercy, but on the reliable exchange condition that once he suffers from accident, others will come to help him. Philanthropy also advocates mutual help, but it does not guarantee as an insurance contract with clear contract terms and legal protection. Of course, philanthropy can save some transaction cost such as the cost of insurance companies and insurance contracts. It is precisely why humans do not rely solely on the insurance industry (the market) but also philanthropy (morality).

As mentioned in previous section, the perquisite of the law of large numbers is that the test condition is unchanged. In reality, different person must have different constraint, so the insurance company will investigate one by one to know the relevant information and charge different premium rate.

What is more, as mentioned before, people will choose to increase information cost as long as it is beneficial to him. In the insurance industry, the behavior of concealing information is more common and serious than in other industries. It is why the "principle of utmost good faith" is one of the four basic principles in the insurance law. As long as one of the parties is found to have

violated this principle, the counterparty can declare the insurance contract invalid.

Section 7: Adverse Selection and Moral Hazard

There is often information asymmetry between buyers and sellers that sellers seem to be smarter than buyers just because sellers know more than buyers. However, it is the opposite in the insurance industry. The insured (buyer of insurance) usually knows the risk status of the subject-matter insured better than the insurance company (seller of insurance), because he is the holder of the subject-matter insured. By contrast, in ordinary commodity trading, the seller holds the commodity before it is sold.

What is more troublesome is that the insured not only knows the risk status of the subject-matter insured better than the insurance company before the insurance contract is concluded, but also still holds it after that, and how he treats it may cause significant changes in the risk status. If the risk status changes, the premium rate (the price of insurance services) should change accordingly, but it is fixed when the contract is concluded. It is difficult for the insurance company to know the information of the changes, and more difficult to negotiate with the insured to change the premium rate even it has known the relevant information, which all means there is high transaction cost.

According to whether the risk (information cost) is before or after the insurance contract is concluded, it can be divided into two categories: adverse selection and moral hazard.

Adverse selection

Adverse selection refers that those with higher risk are more inclined to purchase insurance than those with lower risk, which causes the risk borne by the insurance company is actually higher on average than when all purchase insurance, so the premium rate calculated based on the average risk will be lower.

However, the insurance company will certainly not be so stupid to accept insurance without distinction and charge an average premium rate, but will certainly investigate the risk status of the insured and charge the premium rate accordingly. Take life insurance as an example, the younger the insured, the lower the premium rate. As the age increases, the premium rate will gradually rise. Those who are older than a certain age will be required to have a physical examination, and the insurance company will determine the premium rate and exception clause according to the results of the physical examination, or may even decide to decline. Of course, there will be definitely more cost than when there is no adverse selection, which reflects implies the information cost. As long as there is no government intervention, insurance companies will always try every means to control the adverse selection of the insured, or reduce the relevant information cost.

In fact, there is adverse selection not only exist in the insurance industry, but also in any fields where information asymmetry is serious. For example, banks all like to lend money to the well-run firms with low credit risk. The poor-run firms are very short of money and most in need of bank loans, so it is such high-risk firms that often come to borrow money from banks.

G. A. Akerlof (1940-) is Nobel Prize winner in economics of 2001 for his paper on how adverse selection caused by asymmetric information leads to the shrinkage and disappearance of the used car market. Since buyers do not know the information about the quality of used cars, they will offer a price according to the average quality of the used cars in the market. Those the quality of which is higher than the average level will not accept such a low price and withdraw from the market, which leads to a decline in the average quality of the used cars in the market, so the price

offered by the buyers will also decline. Such a vicious circle will eventually make more and more used cars with a better quality above the average level leave the market, so more and more buyers also leave the market, resulting in shrinkage or even disappearance of the market. Because "lemon" is used in English slang to refer to an artifact (especially an automobile) that is defective or unsatisfactory, Akerlof named it as "lemon market".

However, this Nobel Prize-winning paper attempts to explain a fake phenomenon. In reality, there are not only used car markets everywhere, but also the most prosperous ones are exactly those where the high-quality luxury used cars are for sale. For example, there was a TV program in "Financial and Economical Report of China" in CCTV2 on October 16 in 2010 about the car market in Beijing, including some major used car markets. There are full of luxury cars in the used car markets, even more spectacular than in the new cars markets.

Akerlof is wrong in his extreme ignorance of not only facts but also true economics. Luxury cars with good quality have a much higher rent than those with poor quality, so of course they can afford much more information cost of checking the quality of used cars, which can more effectively eliminate the information asymmetry between buyers and sellers. It is precisely the implication of the law of demand, or more specifically, the implication of the principle of rent matching introduced in Lecture 13: the more luxury cars with higher rent can match the higher information cost paid for better quality inspection of used cars.

(2) Moral hazard

The risk after the conclusion of the insurance contract is moral hazard which refers to the fact that once the insured has the protection of the insurance contract, he will become negligent in preventing the risk. As a result, the probability of the risk is higher than that before the conclusion of insurance contract, which makes the premium rate in the insurance contract be too low.

For example, if one has car without insurance, he will drive very carefully, pay attention to locking and guarding against theft, and try his best to prevent risk as much as possible. However, once insured, since the insurance company is responsible for compensation, one may tend to drive recklessly and not be so careful to guard against the car thieves. It is again the implication of the law of demand. Once the insurance premium is paid, it becomes a historical cost and no longer a constraint to restrict the behaviors of the insured, which means the cost of risk falls, so the insured naturally become inclined not to prevent risk.

In Lecture 3, there is the reading material about how the law on mandatory installation of airbags will objectively encourage drivers to drive recklessly. Some argues that the law on mandatory purchase of the third party liability insurance ¹ for cars will have the same adverse effect. Indeed, if the law stipulates that an insurance company must accept the driver's insurance and is not allowed to adjust the premium rate, it will lead to an increase in the probability of risk. However, as long as it is not the case, the insurance company will take measures to control moral hazard. But how?

There is a clause in an insurance contract that if a traffic accident is settled in one year, the premium rate for the next year will be automatically increased by a certain percentage, even if the insured is not responsible for the accident. Obviously, with this insurance clause, no driver will dare to take chances and engage in moral hazard. The insurance company will raise the premium rate indiscriminately without checking who is responsible for the accident, apparently in order to prevent the driver from bribing the traffic police, thus saves the information cost of investigating the truth as well as the trouble for the insured of bribing the traffic police and deceiving the insurance company. In other words, the transaction costs have been reduced by double due to such seemingly "imparity clause" insurance contract. In the contracts of accidental injury and medical insurance, there is also similar clause complemented with a "reward" clause that the premium rate will be reduced there is no compensation for three consecutive years.

In reality, people will never do nothing with the information cost caused by adverse selection and moral hazard. Instead, they must try their best to reduce the information cost by controlling these two kinds of behaviors. As an economist, he may not be as smart as the competitor in the market, and all he should do is investigate carefully how people reduce costs (including information cost or transaction cost).

By the way, strictly speaking, the insurance based on the law of large numbers mentioned above is "consumption insurance" which means if the specified accident has not happened during the insurance period, the premium paid to the insurance company will not be returned, just like the consumption. Property insurance, personal accident insurance and medical insurance all belong to this kind of insurance. There is also pension insurance belonging to savings insurance which means the policy holders pay premiums when they are young, and the insurance companies return them annuity when they are old. This kind of insurance is actually saving (people deposit money in banks and withdraw it when necessary).

In principle, not like consumption insurance, saving insurance does not compensate the insured whom an accident has happened to, with the premium of the other insured whom no accident has happened to, but pays annuity later with the premium previously paid by the insured. Therefore, if one's insurance compensation is paid by the others' premium in savings insurance, it is similar to stealing money from others, which is strictly forbidden by insurance laws.

1. This insurance means when a traffic accident causes a third party to suffer personal injury, death or direct property damage, the insurance company will pay the economic compensation that the driver is responsible for. ↵

Section 8: The Extremely Low Rate of Return of Social Pension Insurance

When it comes to insurance, the social old-age pension (SOP, a major part of social security, and called “endowment insurance” in China) must be discussed in detail.

From the experience of various nations in the world, there is so serious problem in SOP that it is a terrible fiscal black hole. Sooner or later, it will swallow up a nation’s fiscal money and even its wealth until it goes bankrupt! So far, no nation can solve the problem.

On the surface, the problem of SOP lies in the extremely low rate of return. According to the data in the table of the following reading material “Is Social Security a Good Deal?”, the present value¹ of SOP in the United States is negative for all, and those with more education or born later suffer more serious loss.

READING MATERIAL: IS SOCIAL SECURITY A GOOD DEAL?²

The U.S. Social Security system provides old-age pensions to millions of Americans. In contrast with private pension contracts, participation in Social Security is compulsory. Partly for this reason, there has been continuing debate over almost every aspect of the program.

One question has been whether, viewed as a saving program, Social Security is a “good deal” for participants. That is, viewed as an individual’s saving project, is its Present Value positive? The question is difficult to answer, owing to the many changes that have been made and will surely continue to be made as to required contributions (what percentage of wages has to be paid in Social Security taxes), benefit levels (what pension amounts will be received), and eligibility rules (for example, at what age can a person retire and become a beneficiary).

Looking at the past, it is possible to examine the results for earlier participants. After its initiation in 1935 early enrollees in Social Security did very well. Early participants who had made contributions even for only a very few years received retirement benefits that were calculated much as if they had been under Social Security and contributing to the program all their working lives! (This factor added enormously to the initial political acceptability of the program.) Another initially favorable feature was the large ratio of working contributors to retirees. But these unusually advantageous circumstances were only temporary. By now almost all beneficiaries have been paying into the program all their working lives, as will be the case for new entrants from now on. And the aging of the U.S. population has reduced the fraction of contributors relative to beneficiaries.

Looking toward the future, whether Social Security is a good deal in terms of Present Value depends upon a host of more or less arguable assumptions about what will happen in future years. Although no single evaluation can be expected to resolve debate, a study by Liqun Lee and Andrew J. Rettenmeier offered several interesting results.³

The authors estimated future levels of taxes and benefits and estimated participants’ projected earnings, retirement choices, and longevity. Some illustrative results are shown in the table, for Present Values calculated at a relatively low discount rate of 4%. (As the text indicates, a low discount rate increases the indicated Present Values of investment projects or saving programs.)

For single men the Present Values shown in the table are all negative! The results for single women, and for married men with nonworking wives, are only slightly more favorable: for these groups the Present Values are positive only for workers in the earliest birth years and with the lowest education levels.

The *relatively* more favorable results for those of earlier birth years reflects the initially low Social Security taxes. Rising longevity and the increasingly unfavorable ratio of pensioners to contributors have since caused tax rates to rise sharply. "Progressive" elements have been incorporated from the beginning, to assure a decent minimum standard of benefit even for low-earning workers. So low-income contributors have always done relatively better than high-income contributors. Nevertheless, the table indicates negative Present Values for all the groups tabulated.

COMMENT

Although these results may be questioned in detail, it appears that for most participants Social Security is financially a bad deal. True, Social Security was not intended to be an actuarially sound pension program but has always involved a redistributive element. But even for the least well off participants, the Present Value of Social Security as a savings program is at best only barely positive. Considerations of this kind have motivated proposals for reforms of Social Security that would bring it closer to an actuarially based pension program, while granting participants better control over the placement of their individual contributions.

Projected present values of social security investments – single men

Birth year	High School graduates	College graduates	Graduate school
1940	-27000	-33	-33000
1950	-32000	-45000	-49000
1960	-34000	-53000	-59000
1970	-33000	-58000	-75000
1980	-32000	-63000	-93000

Source: Estimated visually from Lee and Rettenmeier, Figure 2.

It is not an exception unique to the United States. All nations with SOP, without exception, are in deep crisis of severe losses. It is said that many local governments in the United States are facing financial bankruptcy, and the biggest debt comes from the burden of SOP. Similarly, there are several European countries in sovereign debt crisis, but why are the fiscal expenditures so large that they cannot be reduced? One of the main reasons is that the SOP, like a black hole, has swallowed up a large amount of fiscal income. Hong Kong started to implement the "Mandatory Provident Fund Scheme" 10 years ago, which is essentially a SOP. Ten years later, the rate of return is zero! It is actually a good performance compared with the European and American countries mentioned above, but it is only because it has lasted not long enough. As time goes by, the rate of return must fall below zero and the loss will become more and more serious.

Why is the rate of return of SOP so low? The direct reason is administrative monopoly. If any commercial insurance company shows to the customers the table with negative present value in the reading material above, it is impossible for anyone to be so stupid to participate such a pension program. In other words, SOP cannot survive in the market. Now it can still survive just because government forces everyone to participate.

SOP is a saving insurance, so the premium has been being paid several years, more than ten years, or even decades, before the annuity needs to be returned. At the beginning, as long as there are enough new participants, SOP can use the premium paid by the others to return the annuity to those who have soon started to retire. However, as mentioned in last section, saving insurance should return the annuity with the premium paid by the same insured, and should not use the premium of the others as consumption insurance does, because saving insurance is actually not insurance, but saving. A policyholder will retire at the specified time and need the annuity to be returned, which is certain without risk. If one's money in his investment (saving) account is used as the investment (saving) income to another one, it is essentially a Ponzi scheme!

The Ponzi scheme is a kind of "pyramid scheme" and was invented by a businessman named Ponzi in Boston in the early 20th century. Ponzi claimed that he had a complicated investment project that buying some kind of stamps in Europe and then shipping them to the United States for sale. By earning a profit from the difference in the price between Europe and the US, the investors could get a return of up to 50% within 45 days. In order to win people's trust, the first investors all received the promised returns within the specified time, so a large number of investors were attracted. In a short period of one year, as many as 40,000 Boston residents were cheated into giving Ponzi about 15 million US dollars, and each of them "invested" several hundred dollars on average. Before exposed, Ponzi was praised as one of the three greatest Italians along with Columbus and Marconi (the inventor of radio), because he "discovered money" like Columbus discovered the new American continent. When a financial expert exposed Ponzi's investment fraud, he published an article in the newspaper refuting him for knowing nothing.

In fact, Ponzi's scheme is very simple. The investment principal given to him by the new investor is diverted to pay the old investor as the return. It is the logic of a Chinese saying "Knocking down the east wall to build the west wall", which means "robbing Peter to pay Paul". As long as the growth rate of new investment is higher than his promised rate of return on investment, he can always keep this fantastic promise. It is precisely why it is a kind of pyramid scheme: the new investment is the bottom of the pyramid, which is used to support the top of the old investment. However, the law of diminishing marginal returns introduced in Lecture 15 is an iron law that will come into play sooner or later. In other words, the growth rate of new investments will slow down sooner or later, while more and more old investments are requiring returns. Eventually, the bottom of the pyramid will be unable to support the top of it.

Since Ponzi's scheme was invented, it has been used repeatedly, only more beautiful packaged to cover its truth, such as pyramid selling. The latest financial Ponzi's sachem was the case of Bernard Madoff Investment Securities founded by the former Nasdaq chairman, which is the biggest fraud in Wall Street history, causing investors to suffer a loss as much as 50 billion US dollars. In the victims, there were many investment veterans whose IQ is high, and experience is rich, but they were still cheated.

However, the largest and most terrible Ponzi scheme in history is actually the SOPs dominated by governments of various nations. With so low rate of return, it would be impossible for SOP to attract the public to participate. At the beginning, it was accepted because the early insured only needed to pay a few years' premiums to get a pension return much higher than their premiums could support. Where did the extraordinary return on investment come from? It could only come from the premiums paid by those who were young and had to pay for many years before they could receive pension as return. The government has been working hard to expand the coverage of SOP, claiming that it is necessary to cover everyone in the social security to protect him. The truth is, as the number of the insured increases, more and more need to receive pension when they become old and retire. In other words, the top of the pyramid is getting bigger and bigger, so the bottom of it needs to be even bigger and bigger.

Since all are forced to participate SOP, it is actually an administrative monopoly which will not strive to reduce cost and increase return as commercial insurance companies do. However, compulsory insurance is a double-edged sword. On the one hand, it enables the bottom of pyramid to expand rapidly with the strong support of government. On the other hand, it also makes SOP unable to decline those with very high risk, and even unable to charge different premium rate according to different risk. In fact, the rate of return of SOP is extremely low not only because administrative monopoly is always inefficient, and but also because it is unable to adjust the premium rate or even simply decline those with too high risk as commercial insurance companies do, which leads to huge cost or loss.

With the passage of time, the positive side of the double-edged sword of compulsory insurance will become weaker and weaker, because it is impossible for the growth rate of a nation's population in a certain period of time to exceed that of the top of pyramid. The more powerful the government can increase the number of the insured in the early stage, the greater the pressure will be on the growth rate of the bottom to keep up with that of the top in the later stage. It is not necessary for everyone to be covered in SOP that it has come to the end!

As a result, on the one hand, the extraordinary high return at the beginning quickly disappeared or even fell to negative; on the other hand, the government had to inject fiscal funds into SOP to ensure that it would not go bankrupt, and the premium rate (the social security tax) will inevitably increase. Fiscal fund injection is actually worse than the increase in the premium rate. Where does the government get the fiscal fund? It is not a producer, and does not engage in production activities, so it cannot get income from selling products in the market, but from taxes. In other words, the fiscal fund injection also comes from taxpayers, but they are not clearly collected in the form of premium or tax, which makes the insured not know how low the return on his SOP is. If the premium rate is directly increased, the insured knows that the cost of enjoying SOP has increased. If other taxes increase in order to finance the fiscal fund injected to save SOP, he does not know.

1. The term "present value" will be explained in the next lecture of "The Theory of Interest". Here readers only need to know that the higher the rate of return on an investment (saving), the higher the present value, so in principle, no one will choose an investment (saving) with negative present value that means it is a losing investment. [↩](#)

2. This reading material is an excerpt from "Price Theory and Its Application", P472. [↩](#)

3. Liquun Lee and Andrew J. Rettenmeier, "Social Security and Education," National Center for Policy Analysis, Policy Report No. 240 (January 2001). [↩](#)

Section 9: SOP is An Irrelevant Answer

The extremely low rate of return is only the superficial problem of SOP, and the deep defect of it lies in the fact that it is an irrelevant answer to the question it is supposed to answer.

Why do people need SOP? Because the problem of providing for the aged needs to be solved. There has been the problem of providing for the aged since ancient times, so why people did not need SOP until now? In ancient times, people also needed to provide for the aged, which is solved by private means, while SOP attempts to solve it by public means. Why should public means take the place of private means? For thousands of years, private means have solved this problem very well. What are the constraints that have changed now, making private means unable to solve this problem, so that government has to implement SOP?

Let us analyze what is the private means first. There are two means to provide for the aged: one is to save money, which means when one is young, he does not spend all the money earned, but save some, and when he becomes old, he provides for himself by the savings. The other is to bring up children, which means when one is young, he brings up children, and when he becomes old, the children have grown up to earn money to provide for him. The latter is an inter-generational provision for the aged.

Then let us turn to analyze the public means. With careful inspection, there are actually also two means imitating private means: one is save money, which means when one is young, he pays premiums that are essentially savings deposited into SOP, and when he becomes old, he provides for himself by pension withdrawn from his own account. The other is also inter-generational provision, which means government collects premiums (social insurance tax) from those who are working (the younger generation) and transfer them to the retired (the older generation).

SOP has not invented any new means different from private means to solve the problem, so what is the meaning of SOP? Or think in another way: private means, which have worked well for thousands of years and successfully solved the problem of provision for the aged of countless generations, what difficulties can they not resolve now so that SOP is needed to resolve? There are only two difficulties: one is due to the improvement of medical and health care, the lifespan of human has greatly lengthened. When one was young, he expected that he would live to 60 years old (because there is a Chinese saying that "A man seldom lives to be 70 years old"), but when he retires, the lifespan has lengthened to 70 years old or even longer. As a result, the money previously saved according to the lifespan of 60 years old is not enough. The other is inflation. Inflation is rare in ancient times, but it has become so common now.¹ As a result, the money saved when young is eaten away by inflation and is not enough.

Can SOP resolve these two difficulties? Obviously not! Compared with private means, SOP has advantage in two cases: one case is that although some save money prudently, there are still some who squander money casually. The other case is that although some have filial children, there are still some who have no children or have unfilial ones. If there were no SOP, the latter will have no money to provide for the aged and take a miserable life when they are old. With SOP that is compulsory, even those who squander money casually, or have no or unfilial children can at least be provided by SOP when they are old.

However, SOP is supposed not to solve the problem that some squander money casually or have no or unfilial children, but to solve the problem that the purchasing power of savings decreases due to the increase in lifespan and inflation. Some squander money casually or have no or children, which is an unsystematic risk, while the increase in lifespan and the decrease in the purchasing power of savings are systemic risks.

Systematic risk refers to the risk that exists in the whole system, so everyone in it cannot be spared. And unsystematic risk does not exist in the whole system, only some in it will suffer. For example, when one buys stocks, there happens to be a stock market crash. All stocks in the market are falling, so the stocks he buys inevitably also fall, which is a systemic risk. If there is not a stock market crash, only the stocks he mistakenly chooses to buy fall, which is an unsystematic risk. Unsystematic risk stems from personal mistakes, while systematic risk has nothing to do with individuals, and it is a general trend and cannot be dealt with by individuals. Obviously, unsystematic risk does not affect all but only some, so it is not a problem that needs to be worried about and solved by the society or government. By contrast, systematic risk affects all, and cannot be dealt with by individuals, so it needs to be worried about and solved by the society or government.

In a word, SOP is supposed to solve the problem of systematic risk, but it is an irrelevant answer to it by solving the problem of unsystematic risk! The extremely low rate of return of SOP means it does not answer the question correctly. SOP has not solved the problem of systematic risk including the increase in lifespan and the decrease in the purchasing power of savings, but has only solved the problem of unsystematic risk that some do not save money or have no or unfilial children, which means it does not answer the correct question. For the students who do not answer the question correctly, they can be given some credit. For the students who do not answer the correct question, they can be only given a zero.

As mentioned before, if the money in one's investment (savings) account is used to pay another's investment return, the investment (savings) project has essentially become a Ponzi scheme. Although the children's money is also used to provide their parents in private means, it is blood relationship that can clearly define the PPRs of diverting fund. It is quite different from diverting the premium paid by the younger generation to provide the aged of older generation in SOP, because the younger generation does not know how much their money is diverted to the older generation, and they may not be willing to provide for the aged except their own parents.

Some may argue that as long as SOP only applies the mode of self-support that the money in one's social security account can only be used to pay his own pension, and diversion of money is strictly forbidden, it will not become Ponzi's scheme. However, precisely because SOP has failed to solve the problem of systematic risk, it has to continue to face the difficulty that the money (premium) previously saved will be not enough to provide the aged when the insured becomes old. Along with the inefficiency of administrative monopoly, the rate of return of SOP must be negative in the long run, so the money in one's social security account cannot be sufficient to pay the pension promised by government. In order to prevent SOP from going bankrupt, government either diverts the money from the social security accounts of the younger generation, or injects fiscal funds into SOP. The former directly changes SOP from the mode of self-support to inter-generational provision, which makes SOP become Ponzi's scheme. The latter also indirectly does so, because fiscal funds come from taxes, and the major group of taxpayers is, of course, the young generation that is working.

What is more troublesome is that once SOP changes to the mode of inter-generational provision, it will be difficult to abolish it. If government suddenly announces that SOP has an extremely low rate of return, swallows up a large amount of fiscal fund, and is unsustainable, so let it go bankrupt, which is actually a wise decision as a Chinese's saying "to cut one's arm to save one's body". At least it can save the whole society from endlessly continuing to input money into the bottomless black hole. However, how can this decision be approved by the public? The aged who are enjoying the benefits from SOP will not agree, while the young who have been paying to support it and never enjoyed any benefits will neither agree. Although the young may know clearly that SOP will go bankrupt sooner or later, they are of course self-interested in thinking that even if SOP will go bankrupt, it should do so after they have enjoyed it! That is why the SOPs in all nations are dying, but they do not go bankrupt until the day when the last dollar of the national

treasury (even the whole society) is used up. Actually, from the perspective of the whole society, the optimal solution is to let SOP go bankrupt as soon as possible, and return the money left to the public who can go back for private means. It is a proof of the harm of self-interest again.

In fact, one of the true causes of European sovereign debt crisis is not the financial turmoil of 2008, but the huge burden of SOP. As early as before 2008, the governments of many European countries had realized that the bankruptcy of SOP was imminent, but they estimated the time of bankruptcy would be in the middle of the 21st century. Unexpectedly, the financial turmoil of 2008 made the governments spend more money to cope with the crisis, which greatly sped up the progress of the bankruptcy of SOP. The French administration passed a bill to postpone the retirement age, which only delays the time bomb instead of disassembling it. In spite of this, the passage of the bill took a lot of time and caused strong opposition from the public. When the treasury is completely empty, and no matter how the public cries or scolds, the government is still unable to take money out of thin air. Then it is the end of SOP.

As mentioned before, the systematic risk cannot be solved by individuals. However, in reality, private means can deal with the two systematic risks to some extent. As for the increase in lifespan, do not forget that the improvement in medical and health care not only prolongs human's lifespan, but also prolongs the time when one is healthy and fit to continue working. Since lifespan is prolonged and there is not enough money, one can continue to work to earn more money. As long as there is no mandatory retirement age, people will make appropriate choices to solve this problem. It is precisely the implementation of SOP that encourages the aged who are still able to work and live on their own to choose to retire waiting for the provision, and prevents people from making appropriate responses to the change in lifespan.

As for inflation, it is the responsibility of the government, because inflation is caused by it. However, the solution to inflation is not SOP, an irrelevant answer, but monetary policy that controls the money supply, which will be explained in the later lecture of "Monetary Theory". In fact, individuals can also do something to deal with the problem of inflation, such as savings on the assets (like real estate) with floating revenue and against inflation.

Even as for the unsystematic risks that can be effectively solved by SOP, it is better not to use such coercive means. Since the ancient times, there has been morality that advocates one should be thrifty and filial to parents, which is how a society solves such unsystematic risks. And private charity will also be used to take care of those who have no money or children to provide for the aged. Although morality and charity are not as powerful as SOP in solving these unsystematic risks, neither have they negative effects as SOP.

What are the negative effects of SOP? It is not only a huge fiscal burden for a nation as has been analyzed above, but also seriously corrupts the traditional morality. When there is SOP, one will become less willing to save money and keep a good relationship with his parents and children, because the value of doing so decreases, which is only the implication of the law of demand. It is well-known that the westerners do not like to save money as the easterners, and family values, especially the moral values of filial piety, are much weaker in western countries. But it is not well-known that the long-term implementation of SOP is one of the main culprits. Indeed, the SOP has only been implemented in China for a few years, there have been phenomena that it has begun to shake and weaken the traditional morality and family relationship in China.

1. The reason to it will be explained in the later lecture of "Monetary Theory". ↩

Lecture 18: The Theory of Interest

Until now, the factor of **TIME** has not been discussed. In this lecture, the factor of time will be taken into account. In reality, time is valuable, and the price of it "Interest". In some economics textbooks the theory of interest is called "economics of time".

Section 1: The Causes of Interest

The theory of interest or economics of time is the theoretical foundation of finance and investment. Information cost explained in Lecture 17 is used to measure risk, except for which, the rest part of finance and investment are the implication of the theory of interest that was mainly founded by Irving Fisher (1867-1947).

One not only needs to choose different goods to consume at the same time, but also needs to choose to consume the same good at different times. In fact, the same good at different time is already a different good.

Suppose there are options: one is 100 units of grain now, and the other is 100 units of grain a year later, which will one choose? Even if there is no inflation, one will choose 100 units of grain now. Why? There are two reasons: firstly, it sounds bad, but one year later one may be dead, so the 100 units of grain have become meaningless. Only when one is alive can he enjoy any wealth, so one will think that the present value of a good is higher than the future value of it. Of course, since one may not be dead one year later, he is also willing to wait for one year if the future value increases to a certain level such as 110 units of grain one year later. The more 10 units of grain is the interest, which is the price in exchange for one's patiently waiting for one year. It is the cause of interest from the perspective of consumption.

There is also the cause of interest from the perspective of production. If one is given 100 units of grain now, he can sow them as seeds in the field. One year later, there will be more than 100 units of grain as harvest. In other words, the value of goods will naturally increase with the passage of time, and the increasing value is interest.

In conclusion, due to the causes of consumption and production, the present value and future value of a good are different, and the difference between them is interest, while the future value divided by the present value minus 1 is Interest Rate. If one is required to give up the present goods in exchange for the future goods, he must be paid a price that is interest rate because in price theory, price refers to the relative price.

Section 2: Interest Rate

In reality, there are various kinds of interest rates. When one walks into a bank, he will find there are all kinds of interest rates: for demand deposits including installment savings and call deposits, and for time deposits, the durations of which include 3 months, 6 months, 1 year, 2 years and 5 years. These are interest rates for deposit. If one walks into a bank in order to borrow money, there will be even more interest rates that are different for different borrowers (individuals or firms) and for different loan projects. If one does not borrow money through banks but private lending, the interest rate is usually higher.

There are so many different interest rates because they are not pure interest rates, but mixed with other factors which fall into two categories: one is transaction cost, and the other is inflation.

The factor of inflation will be explained in the later lecture of "Monetary Theory". Here is only a brief conclusion: when there is money in an economy, there will be two kinds of value for many variables, Real Value and Nominal Value. The nominal value is the value expressed in money. If there is inflation or deflation, it is different from the real value. The interest rates observable in reality are nominal, and are also called monetary interest rates. The relationship between the nominal value and real value is roughly "nominal value = real value + expected inflation rate". For example, if there is expectation of inflation, the long-term interest rate will be higher than the short-term interest rate, because the future currency will depreciate relative to the current one and needs to be compensated with more interest.

However, if there is only the factor of inflation, when deflation is expected, the long-term interest rate should be lower than the short-term one, because the future currency will appreciate relative to the current one and needs to be subsidized back with less interest. It is much more usually that the long-term interest rate is higher than the short-term one, because there is also the factor of transaction cost.

The factor of transaction cost affecting interest rate is usually risk (measured by information cost that is a kind of transaction cost). The longer the money is lent out, the more difficult it is to accurately predict the future, so the long-term interest rate is higher than the short-term one.

In addition, it is also the reason why the lending rate is often higher than the interest rate for deposit. The interest rate for deposit is used when individuals or firms lend money to a bank, and the credit of a bank is relatively reliable, which means the risk of default or even bankruptcy of a bank is relatively low, so the interest rate it borrows money is relatively low. By contrast, the lending rate is used when a bank lends money to individuals or firms, and the credit of individuals or firms is relatively unreliable, so the interest rate they borrow money is relatively high. A bank's income mainly comes from the difference between the interest rate for deposit and loan, and what it actually earns is transaction cost.

Furthermore, the interest rate of private lending, especially usury, is usually much higher than lending rate of a bank. It is more difficult for individuals or firms repudiate a debt to banks, and laws also protect the right of banks to recover debts. Private lending is often used by borrowers with so poor credit or high risk that even banks are unwilling to lend money to them. In order to recover debts, illegal acts such as violence and gangdom that are not supported by laws, are often used, which implies high transaction cost.

Therefore, the truth of investment is that high rate of return must come from high risk. The higher rate of return (interest) is actually risk premium that is the price paid to those who are willing to bear higher risk. In other words, it is to pay higher transaction cost. All investment frauds make use of people's greed to make them believe that it is a high-yield but not high-risk investment. As the saying says, "Money doesn't grow on trees". Only a fraud will seem to be out of line with it.

It is actually the application of the zero-profit theorem introduced in Lecture 15. According to zero-profit theorem, when all factors including rent and risk are taken into account, the profit for all producers in all kinds of industries must be the same or there can be no supernormal profit, because any industry with supernormal profit will attract outsiders to compete to enter, and competition will dissipate the income higher than cost (rent).

The same profit for all producers in all kinds of industries is precisely the interest rate. No matter what industry a producer is engaged in, he must input a sum of money, either from his own savings or borrowed from others. In the former case, the producer has given up the interest of depositing the money in bank, which is the opportunity cost of the money. In the latter case, the producer directly has to pay back interest as cost. Therefore, even if there is no inflation and risk, the profit margin must be enough to make up for the pure interest rate, or else people will choose to simply deposit the money in banks. If the profit margin is higher than the pure interest rate in an industry, people will borrow money from banks to invest (input) in that industry. Competition will make sure that the profit margin of all kinds of industries be the same as the pure interest rate.

In fact, it is the rate of return on investment of a society. The pure interest rate only clearly represents this rate of return in the form of market price. In other words, the pure interest rate is not determined subjectively by government, but objectively by a nation's production efficiency (the ability of goods to appreciate in production over a period of time). If government directly controls interest rate, or uses monetary policy to influence it, it is similar to price control. As has been analyzed in Lecture 9, price control cannot really change the price of a good, but only mislead people to know the true price, which means increasing the information cost. In reality, due to the factor of inflation and risk, it is not easy for people to know the pure interest rate, which means there is cost for the information about the rate of return on investment by observing interest rate. If government interferes with the pure interest rate to make it more unreliable to represent the return on investment, people will be only misled more seriously to make wrong decisions. But government often blames on so-called "market failure".

The closest to the risk-free pure interest rate is the rate for overnight loans in the interbank market, at which banks lend money to each other overnight. Why do banks do so? Because every day some deposit money into banks and some borrow or withdraw money from banks. The amount of money into and out of banks cannot be exactly the same, and the difference is called "Position". Some banks are shortage of funds when they have less money in and more money out, which is called "short position". They need to borrow money from other banks to make up for it, so they are the borrowers in the interbank market. Some banks, on the other hand, have more money in and less money out, and have surplus funds, which is called "long position". If they do not lend out the surplus money, they will lose the interest, which is an opportunity cost, so they are the lenders in the interbank market.

The loan term for the overnight loans in the interbank market is very short, and all parties involved are banks, so the risk is very low (the transaction cost is very low), so the rate for it is the closest to the pure interest rate. In fact, many other interest rates are charged based on the rate for overnight loans in the interbank market by adding different risk premiums, and are also adjusted directly or indirectly along with it.

The most important rates for overnight loans in the interbank market are the Federal Funds Rate of US and LIBOR (London Interbank Offered Rate) of UK. As for China, it is SHIBOR (Shanghai Interbank Offered Rate). Although there is interest rate control in China, but the interbank market is not directly interfered with interest rate control, so the rate in it should be the most reliable indicator close to the pure interest rate.

Of course, the rates for overnight loans in the interbank market are always fluctuating. In theory, a nation's rate of return on investment will not fluctuate so frequently, but interest rate is also affected by the supply of money from central bank and the short-term changes in demand for money from the banking system and the entire financial system. The former is the major factor that interferes with rate for overnight loans in the interbank market to correctly reflect the pure interest rate. Since Greenspan was in charge of the Federal Reserve System (the central bank of US, Fed for short), the Fed has been very fond of implementing monetary policy by adjusting the federal funds rate, which is precisely one of the main reasons of the subprime crisis in 2008.

In addition, the scandal of LIBOR manipulation by large multinational financial institutions such as UBS (United Bank of Switzerland) has also been reported in European banking, which is price-making (as introduced in Lecture 17) in finance. Since information cost is one of the main causes of so many different interest rates in reality, of course the insiders will make use of the belief that LIBOR is the closest to pure interest rate to make price.

Section 3: The Discount Formula

Based on the concept of interest, a number of relevant concepts, including income, wealth, annuity income, assets, capital and investment, will be introduced in this section.

Let us begin with the concept of "Income". The concept of cost and rent introduced in the previous lectures are all related to income, but this concept has not been introduced, because the nature of it can be explained clearly only in the theory of interest. As mentioned at the very beginning of this lecture, the theory of interest was basically founded by Fisher who had a famous saying about income, "Income is a series of events." What does it mean?

In section 1, there is an example. Suppose one is given 100 units of grain now, he can sow them as seeds in the field. One year later, there will be more than 100 units of grain as harvest.

Attention: he does not obtain the income one year later at the moment of harvest. During the process of the growth of the grains, every day and even every moment there is income that is always changing and different from day to day, sometimes may even be negative such as the wind blowing and the rain hitting the leaves, flowers or ears of corn. Thus, income is a series of events.

There is a way to categorize magnitude by time interval or time point in statistics: flow vs. stock. For example, if water is poured into a pool through a pipe, the amount of water flowing into the pool during a period of time is a flow, while the amount of water in the pool at a certain time-point is a stock. There are some key points about flow and stock.

(1) Different flows can be summed up. For example, if during one hour there is 1 cubic meter of water flowing in, during two hours there will be 2 cubic meters.

(2) Different stocks cannot be summed up. For example, if the stock of water in the pool at 9 o'clock is 1 cubic meter and that at 10 o'clock is 2 cubic meters, these two amounts cannot be added up.

(3) The difference of two stocks is the flow during the period between the two corresponding time-points. For example, if the stock of water in the pool at 10 o'clock is 2 cubic meters and that at 9 o'clock is 1 cubic meter, the difference of them is 1 cubic meter that is the flow from 9 to 10 o'clock.

Income is a series of events, so income is a flow that is not achieved at a certain moment, but formed by continuous events over a period of time.

By contrast, "Wealth" is a stock, but it is not the simple summation of all incomes, because as mentioned in section 1, the same good at different time is actually different goods, and cannot be directly summed up. The income of different periods must be converted into the income of the same period before they are summed up. There are two kinds of conversion: one is to convert the present income (value) into the future income (value), the other is to convert the future income (value) into the present income (value), which is called "discount" and is more usually used.

Suppose there is RMB100 now, which is the present value, what is the future value of it one year later? Suppose the interest rate of one year is 10%, so if this RMB100 is deposited in the bank, and one year later it will become $100 \times (1+10\%) = 110$. Thus, the future value of RMB100 one year later is RMB110 (with an interest rate of 10%).

How about two years later? Here we should pay attention to "compound interest" which means when the RMB110 is deposited in the bank for another more year, not only the principal of RMB100 will receive interest, but the interest of the first year (RMB10) will also receive interest, so the future value two years later is $110 \times (1+10\%)^2 = 121$.

To generalize the above example, the formula for future value is $FV = PV(1 + i)^t$, where FV is the future value, PV is the present value, i is the interest rate, and t is the times of compound interests. Attention: in reality, the compound interest is not necessarily annual. For example, in usury, compound interest is usually monthly, so t is 12 for a year! The borrowers of usury are burdened with heavy interest, and there are the cases that the interest is even higher than the principal, not only because of the high interest rate, but also because t is so large that the interest becomes more and more rapidly like a snowball.

Compound interest is destructive because it leads to exponential growth in interest. There is a story that an Indian chess player played with the king and won. The king asked him what reward he required. He asked for the grains filling the cells of the chessboard as followed: 2 grains were put in the first cell, and then 4 grains in the second cell, and then 16 grains in the third cell...in other words, the number of grains in each cell should be quadratic of the number in previous cell, until all 64 cells were filled. The king thought it was only a small number, but the number of grains increased exponentially, so finally even he had used up all the grains in his country, he still could not meet the requirement of the player.

Let us reverse the above question: suppose there will be RMB100 one year later, which is the future value, what is the present value of it? Suppose the interest rate of one year is 10%, so if a sum of money (PV) is deposited in the bank, and one year later it will become 100, which means $PV (1+10\%) = 100$. Thus, the present value of RMB100 is $PV = \frac{100}{1+10\%} = 90.91$ (with an interest rate of 10%).

If the future value of two years later is RMB100, it is $PV(1 + 10\%)^2 = 100$, or $PV = \frac{100}{1+10\%}^2 = 82.64$. To generalize the above example, the formula for present value or the discount formula is $PV = \frac{FV}{(1+i)^t}$.

If an investment project lasts over a period of compound interest and there are different incomes in each period, the incomes of each period should be discounted respectively, and then summed up like $\sum_{t=0}^n \frac{FV_t}{(1+i)^t}$. Different investment projects have different risks, so i is also different and should be estimated according to the different constraints. Because it is used to discount the future incomes, it is also called "Discount Rate". The risks of each period may also be different, and i will change accordingly. If there is different i in different period, the discount formula for a certain period should be $\frac{FV_t}{(1+i_1)(1+i_2)\dots(1+i_t)}$. If an investment project lasts infinitely, the n in the formula should be replaced by infinity. It is not an unrealistic assumption. In Great Britain, there is a bond (Consol) that has no date of maturity, which means it pays perpetual interest and never repays the principal. Stocks are typical in never-repaying principals. One of the basic assumptions of general accounting is the "going-concern assumption" which also assumes that a firm will keep running in the future and will not go bankruptcy or liquidation.

"Wealth" refers to the sum of the discounted values of all incomes. As mentioned in Lecture 12, the concept of rent is the income of production factor or the value of asset from the first perspective. At that time, because the factor of time has not been introduced, readers may not realize that there is time dimension for the concept of income. Incomes at different times are not comparable, so they need to be discounted before summed up. In other words, there is time dimension for the concept of rent, just like income, while there is no time dimension for the value of asset or wealth.

The above concept of income, rent value, value of asset and wealth are all those from the perspective of producers. If from the perspective of consumers, the use value of a good is also value. From consuming a good, one also obtains income that is non-monetary income or enjoyment from consumption. Think deeply, producers' monetary income from assets will eventually be used to purchase goods for consumption. Money itself is meaningless, but it can exchange for the enjoyment from consumption, which is the ultimate purpose of earning money. As mentioned before, there is no essential difference between the final consumption good and the production factor, because the former can be regarded as the production factor used to produce human. Here is the reverse perspective: the monetary income of a production factor is ultimately non-monetary income or enjoyment from consumption, so they are the same in essence. Thus, both the value of asset and the use value of a good are values, actually wealth that is the sum of the discounted values of all incomes (including monetary and non-monetary income).

The definition of Capital Asset is that any that can bring in income is an asset. If non-monetary income is also taken into account, all goods are assets, because one can not only obtain monetary income by selling a good, but also obtain non-monetary income by consuming it. A good (an asset) is valuable because it can bring in income, and the price of it is the sum of the discounted value of future income. Why? Here is an example as follows.

Suppose one is holding a 3-year bond with a face value of RMB1000 and can receive a fixed annual interest of RMB100. He has held it for 2 years. Now he wants to sell the bond. What is the price?

Obviously, the buyer is willing to buy this bond because the sum of the discounted value of the income of it one year later is not lower than the price that is the cost he has to pay now. Suppose the interest rate used as the discount rate is 8%, the income for a bondholder one year later is the sum of interest of RMB100 and the principal of RMB1000, so the present value of it is $\frac{1000+100}{1+8\%} = 1018.52$. As long as the price of the bond is not higher than RMB1018.52, someone will buy it. Under the competition, the price of it will eventually be RMB1018.52.

In other words, the sum of the discounted value of future income is the price of an asset or a good. And the price of an asset price and interest rate (discount rate) is reciprocal, which means the higher the interest rate, the lower the asset price, and vice versa.

There were reports that the yield rate of the Treasury bond of Greece rose in the financial news during the European sovereign debt crisis. The yield rate is the interest rate of the Treasury bond, and its rising actually implied the decline in the price of it. Few investors were willing to buy, so the oversupply leads to a decline in the price. What is more, as mentioned in section 2, high yield rate must come from high risk. The increase in the yield rate on the Treasury bond does not mean a rise in the rate of return or the pure interest rate, but a rise in the risk premium added to the pure interest rate. Here the rising risk is the risk of default.

The above example implies a premise that different buyers have the same information, so they use the same discount rate. If different buyers have different information, they will use different discount rates, and the sum of the discounted values of the future income will be different, which leads to different "reasonable" price of the asset. Under competition, those who evaluate the asset a higher price will buy it from those who evaluate it a lower price, but they may be wrong in overestimating it, and be punished by the market in the form of investment losses.

The basic theory of finance or investment is not difficult, and this discount formula is almost the all. Based on the discount formula, there are two criteria for investment decision. One is the "Present-Value Rule", which means one should choose the project with positive present value, and larger the better. The other is the "Rate of Return (ROR) rule". The ROR is sometimes called the Internal Rate of Return. The ROR for any project is defined by setting up an equation in present

value form, but with the discount rate treated as an unknown. The discount rate that makes this present value equal to zero is the ROR. One should choose the project with a ROR not lower than the interest rate in the market (such as the lending rate of bank), and the higher the better.

The difficulty lies in how to determine the discount rate. As mentioned before, the key constraints determining the stock prices are difficult to get to know, and are always changing quickly, which implies there is very high information cost.

Besides the discount rate, the future income of some assets is also uncertain. The above example is about a bond with fixed interest rate, so the annual interest can be determined in advance, but if it is a stock that neither repays the principal nor has the obligation to pay a dividend every year, the future income also needs to be estimated according to the information (such as the company's operation in the future). Different investors have different information, and because there is high information cost that makes the behaviors of price-making are rather common in the financial market, who knows what the "reasonable" price is.

As has been pointed out in Lecture 3, the seemingly irrational behaviors in the stock market (financial market) are actually due to ignorance. Those who make money either rely on having more sufficient and accurate information or are just lucky. Luck is a random factor that makes investors win or lose, and in the long run the net effect will be null. In other words, what those who can make money in the long run (ROR is higher than the pure interest rate) earn must be the rent of information.

For the investments that do not rely heavily on time (such as the investments on real estate or doing business), although there is also high information cost, because the constraints do not change so quickly, those who have done detailed surveys and have rich experiences will have a higher chance of making money, and also earn information rent.

Taking real estate as an example, the price of a house is actually determined by the sum of the discounted values of future rent. What is the difference between selling a house at one time and renting it forever? In the former case, a large sum of money is obtained for once, while in the latter case, a small amount of rent is obtained for each period but lasts forever. The sum of the discounted values (present value) of all future rent in the latter case must be equal to the large sum of money in the former case. If it were lower, the owner of the house will choose to sell the house now. If it were higher, he will choose to hold the house for collecting rents. This kind of "arbitrage" behaviors (buying at a lower price and selling at a higher price to obtain the difference) will make the price of a house tends to be equal to the present value of the rents of it. Attention: the price of a house is determined by the sum of the discounted values of future rents of it, not the current rent. As for the future, expectation must be taken into account. As mentioned in Lecture 8, expectation is intention, and should be transferred to the objective facts that form the expectation. Different people have different expectations because they have different information, so information cost again plays a key role.

Section 4: From Asset to Capital, Annuity, Investment and Saving

The price of an asset is Capital Value, which is the most general concept of capital. Capital is the sum of the discounted values of future incomes, while the wealth (value) is the sum of the discounted values of all incomes. The difference between the two is that the former is not including present income. But “present” is a very flexible concept. The income of this year is present income, while that of next year is future income. The income of today is present income, while that of tomorrow is future income. The income of this moment is present income, while that of the next moment is future income. If the present period is as short as a point of time, the income as a flow has no value and present income is close to zero. Thus, there is no difference between capital and wealth.

The incomes in different periods are different or can even be negative. A self-interested person will maximize income, but the question is which period of income is maximized? According to the theory of interest, the income with time dimension cannot be maximized. Actually, one will maximize the sum of the discounted values of all incomes or wealth.

Suppose the t in the discount formula is infinite, we can multiply the wealth by a fixed interest rate, to get an income that is also fixed for each period and called “Annuity Income”. As introduced in Lecture 17, with SOP, people pay premium when they work and receive annuity on time after retirement. The “annuity” is also derived from this concept. Annuity is fixed, will not be different in different periods or even be negative, so it can be maximized by self-interested persons. In other words, the “income” in the “income maximization” mentioned in Lecture 12 actually refers to “annuity income”.

The fixed interest rate used to calculate annuity can be regarded as an estimate of the average magnitude of different interest rates in different periods. Of course, this “average” is only a concept, not a concept defined by a mathematical formula, such as an arithmetic average or a geometric average. Thus, annuity is also a concept, and it can be used to construct a simplified discount formula (called Fisher’s Wealth Equation) in evaluating an investment project: $W = \frac{Y}{r}$, where Y refers to annuity. Compared with the more complicated formal discount formula introduced in section 3, it enables investors to make decisions quickly. For example, suppose one is going to invest in real estate, and there is a house priced RMB200,000 and located where the normal rent is about RMB1,000 per month. With the risk that the house may be empty for a while without income (that is transaction cost) and the trouble of managing the tenants (that is also transaction cost) are taken into account, it can be estimated reliably that the average annual rent is RMB10,000. The price of RMB200,000 is wealth, and the annual rent of RMB10,000 is annuity, so according to Fisher’s Wealth Equation, ROR (r) is 5%, which is much higher than the interest rate of deposit in bank for one year. Thus it can be concluded that such an investment is a good deal.

Fisher also had another famous saying, “Interest is not a part of income, but the whole of income.” With the understanding that “income” in “income maximization” refers to annuity, this Fisher’s saying is easy to understand, because annuity is wealth multiplied by a fixed interest rate, which is precisely interest.

Since both wealth and asset need to use interest rate in the market as discount rate, there will be no wealth or capital if there is no market. At this time, the change in people’s behaviors should be observed from the perspective of marginal change.

The term “investment” in economics is not different from “saving, because they all mean the behavior of giving up the present consumption in exchange for future consumption. The

remaining part of the present income after deducted the consumption is “saving”, while the remaining part of the present income that is not consumed now but left for future consumption is “investment”. Obviously, they are the same, and just the former is from the perspective of present, while the latter is from the perspective of future.

What is more, one’s savings that are not immediately consumed will be converted into investment by others in the society through banks as the intermediaries (indirect finance) or securities markets including stock market and bond market (direct finance).

More generally, all behaviors can be regarded as investments. For example, sleeping is usually regarded as consumption, but it is also investment. If one chooses to read a few more pages of books instead of going to bed in one evening, he can learn more and may help to increase his future income a little. If one chooses to go to bed earlier instead of watching TV in one evening because there will be important work tomorrow, he can do better job tomorrow and increase his future income. In the former case, sleeping is consumption that is given up for the education investment, while in the latter case, sleeping is investment by giving up the entertainment consumption.

As mentioned in Lecture 12, there is no essential difference between production factors and final consumption goods. Bed is usually regarded as a final consumption good, so sleeping is consumption. However, bed can be also regarded a production factor that produce the labor ability, so sleeping is investment, and is also saving that saves physical strength. In a word, all behaviors are essentially the same, but just observed for different perspectives.

Thus, there is no difference between consumption and production. Consumption, production and investment (saving) are all the same. Grain can be used as food (consumption), as seed for sowing (production or investment), and lent to others as food or seed (savings). A piece of land can be used for growing grains (agricultural land), for building factories (industrial land) and for building houses (residential land). Is there any essential difference between grain and land? No, there are all kinds of uses for both, and people will compare various uses of a good to choose the one with the maximal rent.

We should grasp the essence by observing the superficial phenomena, which is precisely the implication of generalization. From the most general perspective, goods (products or consumption goods), production factors and assets are the same. Consumption, production and investment (savings) are the same. Value (use value or price of a good), rent (income or cost without time dimension), capital (price of an asset) and wealth are also the same!

Section 5: The Separation Theorem

As mentioned in section 1, there are two causes of interest. From the perspective of consumption, interest is determined by people's subjective preference for time. From the perspective of production, interest is determined by a nation's objective production capacity or efficiency. Some may wonder different persons have different time preferences, and how can interest be objective? The answer is: as long as the transaction cost of borrowing and lending is low enough, different people's time preference (subjective interest rate) can be adjusted to be consistent with the objective interest rate in the market through borrowing and lending. It is logically similar to the fact that different people's subjective preference for a good (marginal use value) can be adjusted to be consistent with the objective market price through buying and selling.

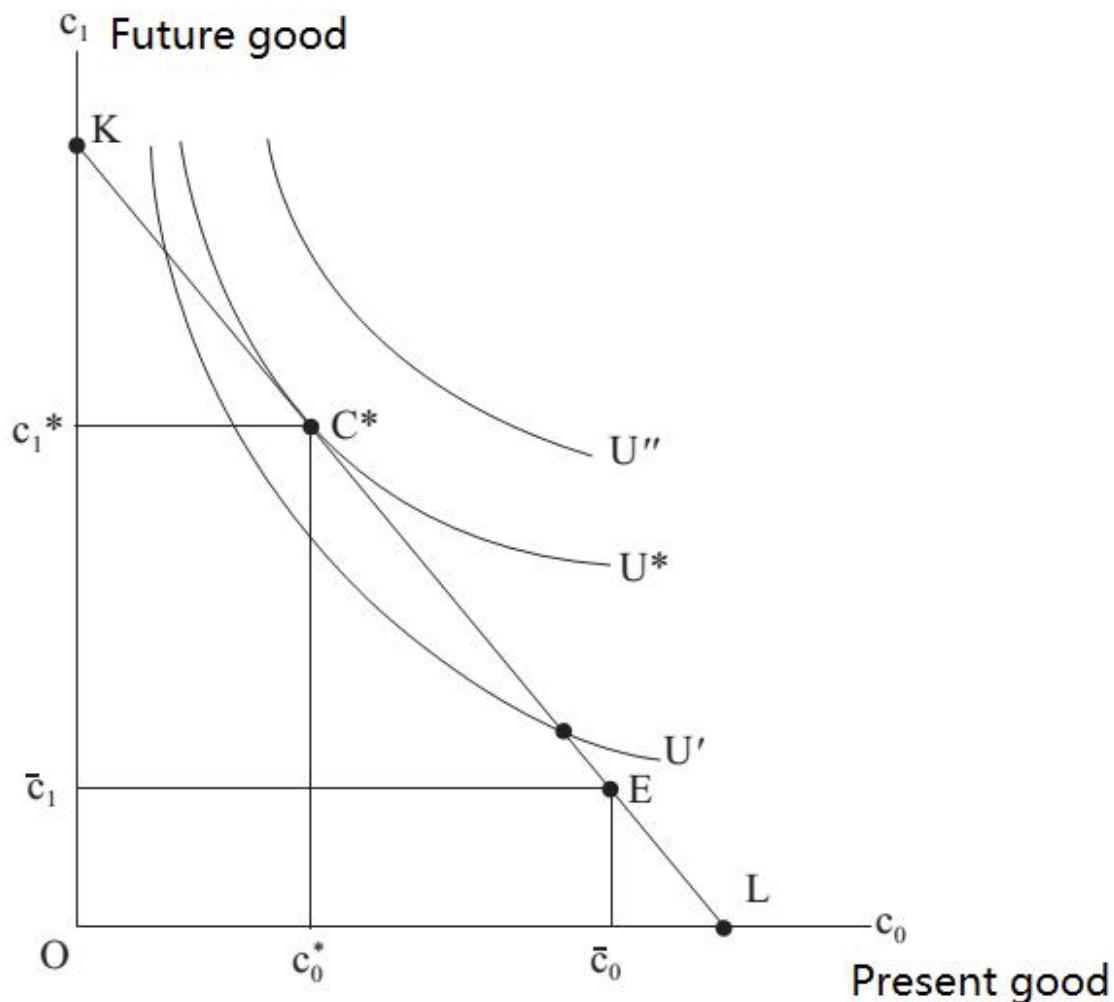


Figure 18-1

Fig.18-1 is logically the same as Fig.10-6 in that the tangent point of the indifference curve and the budget line determines the optimum of consumption, and the horizontal axis is changed to present good and the vertical axis future good. The indifference curve reflects people's preference for time. The slope of the budget line is the interest rate and passes through the initial endowment of wealth (point E). The optimum of consumption is point C*, which means only part of the wealth is consumed now, and the rest is input for production to increase the income available for future consumption. By adjusting the present and future consumption, one can adjust his preference for time to be the same as the interest rate in the market, where is precisely the optimum (equilibrium).

The above is the case when there is no production, and the more general case when there is production is shown in Fig.18-2.

Fig.18-1 is logically the same as Fig.10-6 in that the tangent point of the indifference curve and the budget line determines the optimum of consumption, and the horizontal axis is changed to present good and the vertical axis future good. The indifference curve reflects people's preference for time. The slope of the budget line is the interest rate and passes through the initial endowment of wealth (point E). The optimum of consumption is point C*, which means only part of the wealth is consumed now, and the rest is input for production to increase the income available for future consumption. By adjusting the present and future consumption, one can adjust his preference for time to be the same as the interest rate in the market, where is precisely the optimum (equilibrium).

The above is the case when there is no production, and the more general case when there is production is shown in Fig.18-2.

The curve concave to the origin is called Production Possibility Curve (PPC). Usually, the two axes of PPC represent the output of two different goods, and the curve reflects all the combinations of maximal output that can be achieved when all resources are fully utilized in production. The shape concave to the origin reflects the law of increasing marginal cost (derived from the law of diminishing marginal product), which means along with the increase in output of the good on X-axis, the slope of PPC increases. In other words, the output of the good on Y-axis that must be abandoned also increases, or the opportunity cost to increase the output of good X increases.

In Fig.18-2, the two axes of PPC are replaced by the output of the same good in different periods, thus reflecting all the combinations of maximal output of present good and future good. The price line NN cannot be called budget line here, because it does not pass the point E representing the initial endowment, and does not reflect the income or wealth. The slope of NN only reflects the exchange ratio (relative price) of the X good and Y good, here is the interest rate because it is the exchange ratio of the present good and the future good. The tangent point of NN with PPC is the optimal production point (Q*), while the tangent point of it with the indifference curve U' is the optimal consumption point (C*). As can be seen in Fig.18-2, Q* and C* is not the same point, which means that with the availability of loans the production decision and the consumption decision can be made independently. It is precisely the Separation Theorem.

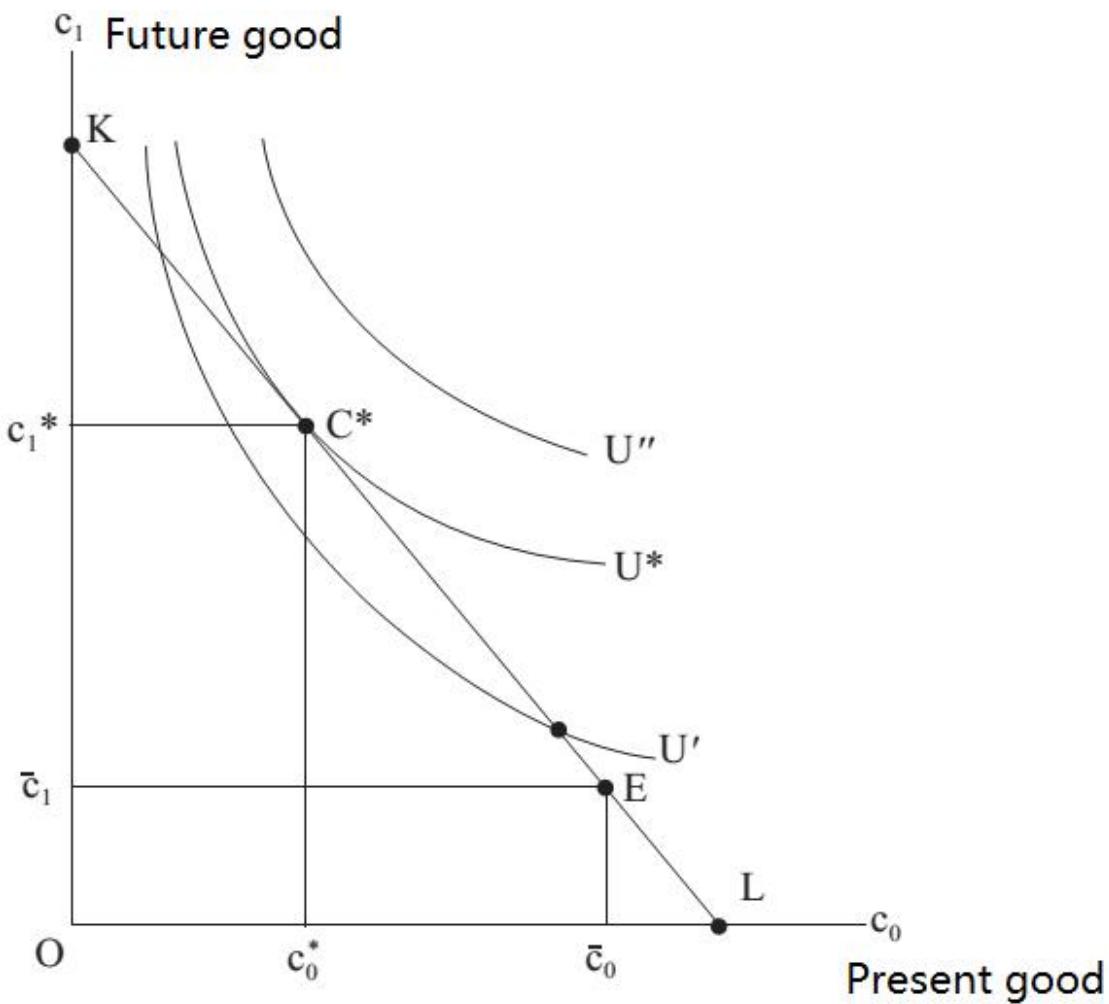


Figure 18-2

There are economic implications of this theorem.

From the individual's perspective, the career choice does not affect consumption pattern although it does affect the income needed for consumption. For example, some like to save money for the future, some like use up all the income without saving, some like to consume the same amount money in different periods, and some like to sometimes consume more and sometimes consume less. All these different consumption patterns do not need different income patterns determined by different careers, as long as the career chosen can bring in the maximal wealth (the sum of the discounted values of all future incomes).

From the social perspective, the production of the whole society has nothing to do with individuals' preferences for time. As long as the constraints of resources, technologies and constitutions are reduced as much as possible to maximize the production capacity or efficiency of the whole society, the social wealth can be maximized. The objective interest rate in the market is determined by the productive capacity or efficiency of the society, not by the subjective preferences for time of individuals.

Of course, there is transaction cost in reality, so requirements of the separation theorem cannot be fully met. For example, from birth to adulthood, a person basically has no income but he still needs to consume. According to the separation theorem, he can borrow money during this period until he can earn money when he works. However, since there is great uncertainty as to how much he can earn from his future work, which means there is very high information cost, usually no one will lend him any money. In order to solve this problem, there is morality that parents are responsible to bring up their children and children should be filial to their parents in human society. Parents with work pay their income for their children's consumption when they are young

and without work or income. And children pay back their income when they grow up and have work and income to support their old and retired parents.

In fact, it is in an implied loan contract where the PPRs of all parties involved are defined by blood ties and promoted by normality, so some of the transaction costs is much less than the explicit loan contract protected by laws. The binding force to ensure the execution of contract is weaker than the explicit one, so it can reduce some, not all, of transaction costs. However, the transaction cost of explicit loan contract is so higher when concluding a bargain that it is hardly possible to come to an agreement on such an explicit contract at all. Although the transaction cost of explicit contract is lower when executing, the overall transaction costs of the implicit contract of mutual support among parents and their children promoted by morality are lower than that of the explicit loan contract between strangers protected by laws. The reason is obvious why it is the former instead of the latter has been chosen to solve the problem that human need to consume much earlier than they can work and have income.

Lecture 19: Monetary Theory

As mentioned in Lecture 6, in price theory, there is no money in the economy, which means the transaction is barter, or it is supposed that the value of money is stable, which means there is no inflation or deflation, so the impact of money on the economy is neutral and does not need to be paid attention to. The monetary theory introduced in this lecture takes the impact of money into account.

Section 1: The Functions of Money

First of all, what are the functions of money? In other words, why do human need money?

As we all know, in ancient times, there was no money in transactions, and people exchanged a good for another good (barter), which is very troublesome, because what one offers is not necessarily what another needs. For example, A has bread and wants to exchange for B's water, while B may not need A's bread but C's pen. Even if A does not need C's pen, he has to exchange his bread for C's pen first, and then exchange the pen for B's water that is what he really needs. However, C may not need A's bread but a book. A has to try to find someone who has the book C needs and at the same time needs his bread...How troublesome it is to match the supply and demand of the transaction!

It is more troublesome that even if B needs A's bread, but he will need it tomorrow, while A has bread today and cannot wait until tomorrow because the bread will perish tomorrow. In other words, in addition to matching the kinds of goods, the transaction also needs to match the time of goods, and some goods are perishable and cannot wait.

What is the solution? The solution that human found in ancient times was money. A good is singled out and can be exchanged for all other goods, and it becomes a universal equivalent or money. Take the example above, A has bread and wants to exchange for B's water, while what B needs is a pen instead of bread. With money, A just tries to find someone who needs his bread and sell it to him in exchange for money, and then A pays the money to buy B's water. Attention: what A needs is in fact not money, but water. But the buyer of the bread does not need to have water, but money. A is willing to accept money instead of water because he knows that B who has water is also willing to accept money. And B is willing to accept money because he also knows that when he pays money to someone who has a pen, that person also accepts money. Everyone accepts money because he knows that all others also accept money. Thus, there is no need to match the supply and demand of the transaction any more, neither does the time of good need to be matched because money will not perish. A will sell out the perishable bread today and exchange for imperishable money, and when he needs something else at any time, he can exchange the money for it.

In conclusion, what are the functions of money?

(1) Money is a medium of exchange that makes it convenient.

(2) Money is a store of value. When one sells a good in exchange for money, the value of the good is temporarily stored in the money. Later, he can exchange the money, actually the value stored in it, for another good.

(3) Money is a ruler of value. In barter, suppose a piece of bread can be exchanged for two bottles of water, which means the value of a piece of bread is equivalent to two bottles of water. What is more, suppose a book can be exchanged for ten pieces of bread. What is the value of the book? There are two versions: one is ten pieces of bread, and the other is twenty bottles of water. There are only three kinds of goods in this example, and there have been two versions of the value of a good. In reality, there are thousands of goods, and how troublesome it can become! If all goods are only exchanged for money, there will be only one version of the values of them. Take the example above, suppose one bottle of water is RMB1, a piece of bread is RMB2, and a book is RMB20, which is the same as the exchange value of them in barter, but now is much more convenient. In other words, money becomes a ruler. Different from the ordinary ruler that is used to measure the length of object, money as a ruler is used to measure the value of good.

Section 2: The Value of Good VS the Value of Money

Since money is a ruler used to measure the value of good, what is the value of good? The value of good is the “use value” introduced in Lecture 10 and the “wealth” introduced in Lecture 18. It comes from the fact that good can be used to make human live and live better. For example, bread can appease hunger and make human live. In addition, bread can be made very delicious, so human take it not only to fill the stomach, but also to enjoy delicious food, which makes human live better. Similarly, a house can shelter human from the wind and rain, which makes human live. A house can be spacious, comfortable and beautiful, which makes human live better.

There is also value of money.

In ancient times, the precious metals such as gold and silver were mainly used as money. There is value of gold or silver as a good. For example, they can be used as bowls or plates, which makes human live. They can also be used as ornaments, which make human live better. Obviously the value of gold or silver as money is quite different.

In modern society, money is only a piece of paper printed specially. If it were not printed as money, the value is extremely low. Once printed as money, the value is equal to the number printed on the paper, much higher than the value of a piece of paper.

It is obvious that the value of money does not come from the use value of the material as an ordinary good, but from the value of the goods it can exchange for. In other words, what human really need is not money, but the goods it can exchange for.

There is such a fable. There was a king who was crazy about gold. One day God gave him a “gold finger”, and anything touched by this gold finger will turn into gold. The king was so happy that he touched everything around him, and turned them all into gold. After a while, his daughter came in, and he hugged her happily. As soon as his gold finger touched her, the living girl became a lifeless gold one. When the king felt hungry, but as soon as he touched the meal, it also turned into gold. At the end of the story, the king starved to death in a room full of gold with grief for his daughter.

The original purpose of this fable is to warn people not to be so greedy, but there is actually profound implication of monetary theory in it. What human really need is not money (whether it is gold or paper money), but those that can be exchanged with the money and can help human live and live better. The real value, or real wealth, comes from the tangible goods like the meal and intangible goods like the love of the daughter in the above fable. Human needs money only because money can exchange for all kinds of goods, which are much more convenient than barter. Therefore, strictly speaking, money itself is worthless, and the value of it comes from what it can exchange for.

The term “financial derivative” sounds very mysterious, and can be only found in countries with highly developed modern finance. In fact, money is the first financial derivative invented by human, and it is as common as air that few realize it. Financial derivatives refer to those whose values derive from real goods. The value of money precisely derives from the goods it exchanges for.

The division of “fictitious economy” and “real economy” seems to be the result of highly developed modern financial system. In fact, once money is invented by human, there has been fictitious economy (monetary economy) in human society. The economy only with barter is a pure real economy. When there is money helping to make exchange convenient, a monetary economy has

been established on the basis of the real economy.

Section 3: The Cause of Inflation

It is said that financial crisis is caused by the disconnection between the fictitious economy and the real economy. It sounds very profound, but in fact it is logically the same as disconnection between the value of money and that of the goods. How can that happen? It is precisely the problem studied by monetary theory.

As mentioned in previous section, the value of money derives from the goods it exchanges for. For example, suppose there are 100 units of goods in a nation, and the government has issued 100 units of money. In other words, there are 100 units of money in this nation to buy 100 units of goods, so the average price of a good is $100 / 100 = 1$, which is called price level, not the specific price of a certain good, but the average price of all goods as a whole. Thus, the money with a face value of 1 can, on average, buy 1 unit of good. The value of 1 unit of money derives from the value of 1 unit of good.

Suppose the 100 units of goods in the above nation have not changed, but the government prints 100 units more of money, which means there are now 200 units of money in the nation. There are now 200 units of money to buy goods that are still 100 units, so the average price of a good is $200 / 100 = 2$, which means the price level has risen. It is precisely inflation! The cause of inflation is so simple. It is the quantity of money is out of line with that of real goods that causes the price level to rise. It is also the essence of the cause of all economic or financial crises: the disconnection between the fictitious economy and the real economy. Now, the value of 1 unit of money drops to $1 / 2$ unit of goods. In other words, the value of money is the purchasing power of it.

Since the value of money actually derives from goods, if the quantity of goods does not increase, neither should the quantity of money increase. Money cannot appreciate by increasing itself, but just depreciates. In other words, internal inflation is the same as external currency devaluation, and both are the decline in the value of money. Internally money is exchanged with domestic goods, while externally money is exchanged directly with foreign currency but indirectly with foreign goods because foreign currency is exchanged with foreign goods.¹ If the value of foreign currency remains unchanged, but the purchasing power of domestic currency declines, it is clear that domestic currency will depreciate relative to the foreign currency.

This truth is easy to understand with common sense. If money itself were valuable and did not depend on the goods it exchanges for, a nation's wealth would be easily increase by just printing more money, and would not need to work hard for production, and all factories and shops could be shut down. How could it be true?

Therefore, in order to stabilize the value of money, the quantity of it must change along with the quantity of goods. For example, if the goods increase by 10%, from the original 100 units to 110 units, the money can also increase from 100 units to 110 units, no more, no less. If it increases more than the goods, there will be inflation. If it increases less than the goods, there will be deflation.

1. Attention: this conclusion is based on the premise that the value of foreign currency remains unchanged. ↩

Section 4: The History of Money

There was rarely inflation in ancient times, but it has become so common in modern economy. Why? As mentioned in the previous section, the cause of inflation is the disconnection between the quantity of money and that of goods. So the answer must be that this disconnection is rather common in modern economy.

But why is the quantity of money easily out of line with that of goods in modern economy? There are three reasons.

(1) In ancient times, money is precious metals such as gold or silver, while in modern economy, money is paper money. The quantity of precious metals will not easily increase, while paper money can be easily printed by the government.

In fact, there were inflations or even super-inflations in ancient times. The following is a brief history of money.

Marx had said, "Gold and silver are not money in nature, but money is gold and silver naturally." Indeed, in the very early days of mankind, money was not precious metals such as gold or silver, but salt and rare shells had been used as money.

According to the functions of money, if only as a medium of exchange, any kind of goods can be money. However, if as a store and a ruler of value, precious metals are the most suitable goods. In order to store value, money itself cannot be perishable or easily broken, resulting in loss of value. Salt and shells are not perishable, but shells are easily broken. Gold is the most chemically stable, which means it is the most difficult to start chemical reactions. For example gold will remain unchanged when burning at high temperature, and will not be eroded by strong acid or alkali. In fact, rot means chemical reactions have changed the original material with higher value into another one with little or even no value.

What is more, if as a ruler of value, the quantity of money cannot increase too easily, or else it will exceed the increase in the quantity of goods, causing inflation. That is why when shells are used as money, they must be rare ones. Gold is also quite suitable in this respect. In fact, the quantity of gold in ancient China was so few that it is silver other than gold that was the main money.

Gold also has the advantage that it is stable chemically, but it is so soft physically that it can be bitten by teeth, which is convenient to cut up gold into small parts for the small transactions. It is why diamond, which is also imperishable and rare but too hard, fails to become money.

Thus, gold and silver (especially gold) are really, as Marx said, born to be money. If as an ordinary good, gold is actually the most useless. It is too stable chemically to be changed into more useful materials through chemical reactions. By contrast, silver will turn into silver sulfide which is black when exposed to sulfides, which is why in ancient times, silver was often made as food vessel to track toxicity. And gold is so soft physically that it cannot be as widely used as iron and steel. Except as money, gold is good for nothing!

Although precious metals were mainly used as money in ancient times, but they were too rare and of too high value for small transactions in daily life, there are also coins of base metals (such as copper, iron and nickel). These coins have face values stipulated by government, and there will be inflation or deflation if the face values are out of line with the contents of the base metals and the price ratio with precious metals. In ancient China, there had been a monetary idea of "the match of mother and children", which meant that the money of precious metal was like the mother, while the coins of base metal were like the children. The par value of them must be in line with their respective content and price ratio. However, some ancient governments still set the

face values of the coins regardless of it, which in fact created inflation by reducing the contents of coins or inflating the face values to increase the money supply. The following are two such examples in history.

One is "Qianyuan Treasures" in Tang Dynasty in ancient China. Qianyuan was the name of years during which an emperor in Tang Dynasty. He faced a fiscal crisis caused by huge military expenditures. He was suggested to create inflation by issuing a new currency historically called "Qianyuan Treasures". Here "treasures" refer to a kind of coins much heavier. Although the coin was very heavy, but in fact the contents of it were not worth the face values. For example, one "Qianyuan Treasure" weighed 5.97g, and was stipulated to be equal to 10 "Kaiyuan Tongbao" that was the old currency and weighed 3.98g. Obviously, according to the relative weight, one "Qianyuan Treasure" should be only equal to about 1.5 "Kaiyuan Tongbao", but the government set the ratio to be 10 times. No wonder the emperor openly declared this issuance of new currency had made the national treasury full and gotten 10 times of profits. Only a year later, he issued another new currency with the weight of 11.94g that was 3 times of "Kaiyuan Tongbao" in weight but 50 times in face value. According to official records, the direct consequence of such crazy issuances of heavy coins is nothing but inflation. The price of grain was soaring, and there were a lot of people who were starved to death along the roads.

The other is the reform of remonetization by Shigehide Ogiwara of Japanese Shogunate. The fifth General of Shogunate also faced a fiscal crisis caused by the profligacy of the previous two generals. Shigehide Ogiwara, the chief treasurer of Shogunate solved this problem by a reform of remonetization which meant retrieving the coins in circulation and reminting. For example, the most widely circulated in the market at that time was "Keicho koban", and after remonetization the weight of it remained the same, but the content of gold decreased from 85.69% to 57.37%. Silver coins were also reminted. For example, the content of silver in "Keicho chogin" decreased from 80% to 64%. At the very beginning, this remonetization did greatly increase the fiscal income of Shogunate. But as a large number of depreciated currencies flowed into the market, it caused serious inflation and discredited the Shogunate.

According to the lessons of Chinese history, the basic pattern of the collapse of most imperial dynasties is like that: the first stage is fiscal crisis caused by profligacy or militarism. To solve the fiscal crisis, the appropriate means should be to increase income and reduce expenditure. However, it is almost impossible for an emperor who has been accustomed to spending lavishly to reduce expenditure. And the increase in tax is only convenient for corrupt officials to deprive people's wealth, so most of the increase in tax just flows into private pockets instead of national treasury. If the tax is continually increased, people cannot afford and may rise in revolt. The right path is not easy to follow, so the government will be tempted to follow the devious way of creating inflation which can easily increase fiscal incomes at the very beginning, while the evil consequence will emerge later, just as a Chinese saying "Quench a thirst with poison". The ancients did not know the monetary theory, so inflation was easily out of control once created, and quickly worsened to be super-inflation, seriously damaging the economy. It is the second stage. At this time, as long as there happens natural disaster, such as flood or drought, hitting the ancient pillar industry of agriculture, and the officials who are responsible for relief seize the opportunity to make a fortune, there must be uprising, which is the third stage. Thus, the trilogy of fiscal crisis, inflation and natural or man-made disasters that cause uprising had played the funeral symphony of most dynasties in ancient China.

In ancient times, because it was necessary to issue or remint a large number of metal money to artificially create inflation, it was much more troublesome to do so. If it were not for the imminent fiscal crisis, and it were difficult to increase income and reduce expenditure by right means, government would seldom do so. By contrast, in modern society, government can easily print paper money with face value much higher than the value of paper as a good. Under the guidance of monetary theory, as long as government does not go too far, there will not be super-inflation,

so the public will only complain and withstand it. No wonder the so-called "moderate" inflation below 10% will become a common phenomenon. However, as mentioned in Lecture 18, compound interest is destructive, so even if the annual inflation rate is not very high, it is still a terrible killer for long-term investments (savings).

(2) Keynes's theory advocates that government should use monetary policy to create inflation to deal with economic depression, which provides a seemingly "scientific" excuse to increase the quantity of money out of line with that of the real economy. By contrast, in ancient times, although government did not know monetary theory, neither did it think to intervene in the market with monetary policy to promote economic growth.

(3) In most modern western or developed nations, there are trade unions and welfare systems that interfere with the market and make the economy lack of the flexibility to adapt to changes, such as the wage rigidity. By inflation, it is possible to temporarily adjust (real) price (wage) and increase flexibility, thus promoting economic growth.

Even Friedman, the leading figure in the monetary school, thinks that moderate inflation is a "lubricant" for the economy, which is actually a compromise policy proposal based on the constraint of the lack of flexibility in the economic system. Inflation has no function of lubricant in essence. It will be explained in detail in the following Lecture 20.

Section 5: The Harm of Inflation and Deflation

As we all know, inflation and deflation is bad. But why? In this section, it will be explained with economics.

As mentioned in section 1, one of the functions of money is the ruler of value. Since it is a ruler, the length of it should not change. Just imagine if an engineer uses a ruler to measure the length of an object, one meter today is so long, but tomorrow it will change, while the real length of the object has not changed, how can the engineer use such an frequently changing ruler to build a house? Logically similarly, if money, as the ruler of value of goods, always changes, while the real value of a good does not change, how can a producer or consumer make correct decisions?

As has explained in Lecture 6, the cause of the Great Depression in the United States is that the Federal Reserve injected too much money into the economy in the 1920s, resulting in inflation that misled people to make wrong production and consumption decisions. In other words, in a world where the value of money is unstable, the change in the observable monetary price is mixed with inflation. People should only respond to the change in supply and demand, but they cannot tell whether there is inflation, so their decisions are more or less wrong. It is precisely the cause of economic instability (economic fluctuation or business cycle): the fictitious economy (such as money) is out of line with the real economy, which is not market failure at all. When the disconnection is very serious, there will be economic crisis, usually preceded by stock market crash, currency crisis (sharp devaluation of the money value), financial crisis, etc.

The previous statement that the value of money derives from that of goods is actually the viewpoint of classical economics. Classical economists think money is only a veil covered on the goods. An increase in the quantity of money will not lead to an increase in the total value of money in the economy, but a decrease in the value of each unit of money, thus causing an increase in the nominal price (monetary price) of goods, or inflation. The real price of goods is relative price and will not change because the supply-demand relationship reflected by it has not changed, so the change in the quantity of money has no impact on the real economy. It is called "monetary neutrality".

However, after the GD in the US, Keynes put forward the opposite view. He advocates that money is not neutral, which means the change in the quantity of money has an impact on the real economy. He believes that government can regulate the economy through monetary policy: increasing the money supply during the depression and creating inflation to promote economic growth, and vice versa.

Classical economists are right about the value of money. However, Keynes's proposition is more reasonable about whether the money is neutral, but his explanation is wrong. The correct explanation is that the theory of supply and demand is based on the premise that there is no money (or money is neutral). In a world without money, people make decisions according to the real price (relative price). In a world with money, because the information cost of monetary price (nominal price) is much lower than that of real price, people actually make decisions based on monetary price, and there may be "money illusion". When monetary price is actually out of line with relative price, people will be misled by it to make wrong decisions, just as illuded by money. It is somewhat similar to the case mentioned in Lecture 11: it is opportunity cost instead of historical cost that determines people's behaviors. However, when the information cost is high, people will estimate the opportunity cost with the historical cost because the information cost of latter is lower, but at the cost of the increasing risk of making mistakes.

Because people make decisions based on monetary price, when it is out of line with real price due to wrong monetary policy, people will be misled by money illusion to make wrong decisions. For example, producers find the (monetary) price of their products rise, not knowing that it is only due to inflation, so they think their products are popular in the market and expand production. As producers expand production, the goods (national income or social wealth) in the real economy have really increased. In other words, the classical economists' view that money is neutral and will not affect the real economy ignores that there is information cost (transaction cost) in a world with money. However, illusion is illusion, and sooner or later people will be disillusioned. Keynes wants to make use of money illusion to achieve the purpose of promoting economic growth by government intervention (increasing money supply to create inflation) in the economy, which is not only effective in the short run¹, but will also eventually result in bitter consequences. The disconnection between monetary economy and real economy to a certain extent will inevitably lead to economic crisis, just as those who wrongly estimate opportunity cost with historical cost will sooner or later face the consequences of losses.

The increase in production due to money illusion will eventually turn into overproduction because producers will drastically shrink production when they are disillusioned, which leads to economic fluctuations. The so-called "Business Cycle" is not caused by the contradiction between socialized mass production and private ownership as Marx thought, nor by the free operation of market that will lead to the total demand less than the total supply as Keynes thought, but by the government's wrong policies, one of which is precisely wrong monetary policy that misleads people by the "noise" of inflation. The other wrong policies include the "evil laws" such as trade unions, minimum wage law, social welfare and trade protectionism which all infringe PPRs and undermine or even simply abolish the price criteria (market mechanism).

The above is to explain how inflation weakens and loses the function of money as a ruler of value, and even misleads economic entities to make wrong decisions. In addition, inflation is the devaluation of money, which harms those with fixed income. Inflation also discourages people's enthusiasm for savings, because if RMB100 this year is saved, the principal and interest next year may not buy what is worth RMB100 this year, and then why will people choose to save instead of consume immediately? As mentioned in section 1, one of the functions of money is a store of value, because money is not perishable like such goods as bread. However, it only means that the physical form of money will not perish, but the value form may decrease due to inflation. If there is inflation of 100%, which means RMB1 now will be only worth RMB0.5 in the future, what is the difference with that half piece of bread will perish tomorrow and the value (price) of it falling from RMB1 to RMB0.5? That goods perish is devaluation, which is essentially the same with money devaluation due to inflation. In a word, not only the function of money as a ruler of value but also that of money as a store of value will be greatly undermined by inflation.

In such an extreme case as super-inflation (the inflation rate is over 100%), the prices are soaring, and may be different in the morning and in the afternoon of the same day. Thus, the function of money as a store of value will entirely lose, and people will exchange the money for goods as soon as they get it.

The last time China experienced such super-inflation was before liberation (1949) when the Kuomintang Party (KMT) issued a new currency called "gold yuan" note. The national government (controlled by KMT) demanded people to exchange gold and silver for this currency, and that was why the new currency was called "gold yuan" note.²

"Gold yuan" notes were issued on August 19, 1948, with a legal gold content of 0.22217 li (a Chinese unit of weight that is equal to 0.05 grams). After the issuance of "gold yuan" notes, people are prohibited from possessing gold, silver and foreign exchange which must be changed into "gold yuan" notes before September 30 of that year, or else will be confiscated. At the very beginning, the national government's central bank announced to plan to issue 2 billion "gold

"yuan" notes. However, by the end of 1948, the issuance of "gold yuan" notes had been more than 8.1 billion, and rise to 5 trillion in April 1949 and 130 trillion in June 1949. The consequence is naturally soaring prices. In May 1949, the price of a dan (a Chinese unit of dry measure for grain that is equal to 60 kilograms) of rice was over 400 million "gold yuan" notes. The face value of the "gold yuan" note keeps setting new records: at first, the largest face value was 1000, and then 10000, 100000, 500000, 1 million and even 5 million! Even with such high face value, it still needs a large bundle or even a big bag of money to buy rice because the price of rice had risen to 400 million "gold yuan" notes.

In June 1949, after successively occupying Nanjing and Shanghai, the People's Liberation Army (PLA) announced to stop the circulation of "gold yuan" notes by exchanging them at a rate of 100,000 to 1 for RMB. A large number of people's savings vanished overnight! My mother told me the story of my grandmother when I was a child: The total amounts of money she had worked all her life to save was not enough for 100,000 "gold yuan" notes, so she could not even exchange them for RMB1. She could only be crying while burning all the "gold yuan" notes to make them play their final values as paper. From then on, she never believed in any paper money! Any time my parents gave her money, she immediately spent them all. She also left us such a "famous saying", "Eat up, use up, and keep healthy!"

These histories written in blood and tears teach us that inflation (especially super-inflation) is a serious blow to savings. Investments come from savings. Without savings, how can there be investments? Without investments, how can there be productions? Without productions, how can the demand of consumption be met, let alone economic growth? In fact, in the developing countries in South America and Africa, the rate of savings is not as high as that in East Asia. The usual explanation is that people of East Asia like to save because of the influence of traditional Chinese ideas. However, this explanation based on the unobservable ideas is irrefutable, so it is not scientific. The scientific explanation is that the inflation rate in South American and African countries is usually much higher than that in East Asian countries, which of course greatly weakens people's enthusiasm for saving.

It can be observed in the history of "gold yuan" note that even the currency like "gold yuan" note is accepted by people, but is just exchanged quickly for goods. It means that even in the period of super-inflation, the two major functions of money as a ruler and a store of value have lost, but the function as a medium of exchange still plays a role, because it can decrease the transaction cost of barter so much that people would rather accept the rapidly depreciated money than barter. However, if "gold yuan" note had continued to depreciate, there would be a day when it was rejected completely, and then the function of money as a medium of exchange would be also lost. Without money to facilitate the transactions, the trouble (transaction cost) will increase greatly. In modern society, producers do not produce for their own consumption, but to sell their products to consumers. If transaction cost is so high that transactions greatly decrease, so will the production do. Without productions, how can there be economic growth?

It is precisely why the stability of monetary value is so important! Without the stability of monetary value, the market's function of automatically adjusting the behaviors of various economic entities with price will be lost, while the three major functions of money (a medium of exchange, a store of value and a ruler of value) will also be lost one after another, resulting in increase in transaction cost. The end may be economic crisis, and even the collapse of the political regime!

Monetary system is an institution. As pointed out in Lecture 14, all kinds of institutions of human are created to reduce transaction cost, but inevitably they also bring about an increase in transaction cost in some respects. The three major functions of money are precisely the positive effects of monetary system in reducing transaction cost, while the instability of monetary value is the negative effect of it in increasing transaction cost. In order to make the monetary system play

a positive role as much as possible with as little as possible negative effect, the stability of the monetary value that means the quantity of money should change along with that of the quantity of goods becomes the core issue to a monetary system. In other words, the monetary neutrality advocated by classical economics should be understood not as an equilibrium that can be reached naturally, but as a desirable goal of monetary policy of government (central bank).

1. Of course, Keynes has a famous saying, "In the long run, we will all die." [↳](#)

2. The term "yuan" is the most common monetary unit in China, meaning "dollar". [↳](#)

Section 6: The History of Paper Money

Since the value of money derives from that of goods, in order to stabilize the value (purchasing power) of money, it is necessary to make the quantity of money change along with that of the quantity of goods in real economy.

In reality, there are two major difficulties in implementing this principle. Firstly, the issuer does not want to implement this principle at all, but only print money at will. Secondly, it is difficult to know clearly what the effective quantity of money in an economy is, which is the problem of measuring the money supply.

Let us discuss the first difficulty, and start with introducing the history of paper money and the evolution of the international monetary system.

As mentioned in the introduction about the history of money in previous section, human had used salt and shells as money, but later they all turned to precious metals such as gold that is most suitable for money. In order to facilitate small transactions, coins of base metals had been issued. Coins are subsidiary, while gold (or silver) is the standard currency, because the value of coins is determined by the price ratio of the base metals and precious metals contained in them. In other words, the monetary system of ancient societies was gold standard (that in China was silver standard). Within a nation, coins can be used in transaction, but transactions between different nations (international trade) are settled in gold. The ratio between coins in different nations is fixed, which is called fixed exchange rate system. For example, if 10 pounds in Britain is equivalent to 1 ounce of gold, while RMB20 in China is equivalent to 1 ounce of gold, 1 pound in Britain is equivalent to RMB2 in China, and this exchange rate between Britain and China is fixed.

However, whether it is the standard currency of precious metals or the auxiliary coins of base metals, they are heavy and inconvenient to carry and transport. With the development of long-distance trade, this inconvenience becomes very troublesome. As early as the Song Dynasty, "Jiaozi" that is actually a receipt appeared in China. A businessman deposited his gold in a money house (the predecessor of bank), and the money house would give him a receipt stating when and how much gold was deposited by whom. The businessman took the receipt to another city, showed it to the branch of the same money house in that city, and withdrew the amount of gold recorded on the receipt. Obviously, the money house was responsible for keeping the businessmen's gold, and the businessman enjoyed the service of the withdrawal of deposits from different places provided by bank today. The businessman with Jiaozi can avoid the trouble of carrying and transporting heavy gold and reduce the risk of being robbed of gold during his trip.

With the development of trade, some businessmen, when they wanted to pay for their goods, simply transferred their Jiaozi to the sellers who took them to the bank to withdraw gold. Since the name of the buyer instead of the seller was written on the receipt, the buyer needed to declare on the back of the Jiaozi, "I would like to transfer the withdrawal right of the Jiaozi to so and so." It is called "endorsement". Until today, commercial bills can be transferred by such endorsement. The seller took the receipt with endorsement to the money house, and the money house would give him the gold deposited by the buyer according to the amount written on the Jiaozi.

Afterward, the Jiaozi issued by the creditable money houses was widely accepted, and the sellers were no longer in a hurry to withdraw the gold because they believed that the creditable money houses did have so much gold in their warehouses. Instead, they held the Jiaozi in their hands, and when they wanted to buy some goods, they could transfer them to the sellers by "re-endorsement", thus reducing the trouble of drawing heavy gold for payment.

However, with such continuous endorsements, the space behind the Jiaozi was limited and could not contain so many endorsements soon. At last, the credible money houses issued formatted Jiaozi with the same face value. For example, each Jiaozi had the face value of 1 liang (a Chinese unit of weight) of gold, so the businessman who deposits 10 liang of gold would be given 10 such Jiaozi that are bearer notes and did not need endorsements when transferring. The money houses would pay the gold according to the face value to anyone who presented the Jiaozi. Thus, such Jiaozi had already become paper money.

Strictly speaking, this kind of paper money was actually bank note issued by private money houses or banks. In English, the words such as "note" and "bill" can refer to money, which implies that the original paper money was bank note.

These bank notes were backed by sufficient gold reserves, not credit money or Fiat Money. Because a bank gave the paper money to a businessman according to how much gold he had deposited in the bank, there were gold reserves equivalent to the face value of each paper money in the bank's warehouse. This kind of paper money is called "paper money with full reserves". The bank notes with sufficient reserves are actually tokens (not the same as coins made of base metal), which means they are only the representatives of the standard currency such as gold.

School dining halls usually required students to exchange RMB for meal tickets before. The exchange rate is 1:1, which means 1RMB exchanged for 1 meal ticket. This meal ticket was precisely a token, which was the representative of RMB to circulate only in the dining hall. It was logically the same that bank notes could only circulate within a nation. Another example is the round and plastic tokens used in game rooms.

Token is characterized in that it is strictly issued in accordance with the currency it represents. This token-based paper money system can be called the "true standard system". If gold is the standard currency, it is the "true gold standard system", which is not much different from the "gold standard system" where gold as money directly circulates in the economy. In the true gold standard system, gold is still the standard currency, and paper money just represents it to circulate. The money supply is determined by the quantity of gold in circulation and in inventory, but the circulation of paper currency greatly reduces the transaction cost. Since bank notes are issued strictly according to the gold received, it is impossible to be overprinting and currency crisis in true gold standard system. As long as the quantity of gold does not suddenly increase sharply, which is far beyond the increase in the quantity of goods in real economy, it is also impossible to be serious inflation.

However, the problem is that once paper money is accepted and circulated in the economy, the bank will be tempted issue paper money out of line with the gold reserves, thus actually violating the true gold standard and becoming the pseudo gold standard. The cost of printing a piece of paper money with a face value of RMB100 may only be 0.01RMB, while those who want to get it from the bank must pay for the goods worth RMB100. If a bank quietly issues a piece of paper money with a face value of RMB100, the profit is RMB99.99! This extremely high profit is called "seigniorage".

Under the temptation of such high profit, it is no wonder that a bank will tend to print paper money out of line with the gold reserves. The consequence, of course, is inflation. At the very beginning, it is difficult to find that the bank does so, but eventually people will find something is wrong and begin to doubt why the bank issues so many notes that are obviously more than the gold reserves. Once they do not believe the bank can exchange all notes it issues for gold, people will run on the bank. The bank issues more paper money than the gold reserves, and certainly does not have so much gold to meet the need of cashing, so can only shut down and absconded with the money. It is financial fraud from the micro perspective, and a financial crisis from the macro perspective.

In Marx's time, many countries did not have central banks, and neither did they have monetary policies. However, there had been economic crises so frequently (although far less serious than the Great Depression of 1929) that Marx doubted there was inherent defect of capitalism (market economy). In fact, the problem did not lie in capitalism, but in the barbaric and chaotic issuance of paper money. As has been clearly analyzed in previous section, in a world with money, instability in the value of money will lead to inflation or deflation, which will mislead economic entities to make wrong decisions and eventually result in economic fluctuations (business cycles). In essence, it is the increase in transaction cost of monetary system. In modern economy, inflation is mainly caused by the abuse of monetary policy of government (central bank), while in Marx's time, inflation is mainly caused by the lack of management on the issuance of paper money because any private bank could issue paper money at will.

The Austrian school in economics advocates completely free market and opposes any form of government regulations, which is equivalent to "anarchism" in political philosophy. However, as has pointed out in Lecture 5, market system can reduce some transaction costs, but it will also bring another transaction costs. If the transaction costs reduced the market are less than those it increases under certain constraints, self-interested human will not choose market. If market can always reduce transaction costs more than those it increases, there will be no governments, firms, and other organizations, nor will be morality and a lot of non-price criteria voluntarily chosen by human in the world.

Monetary system is a typical example. If the right to issue paper money is delineated to any private banks, tempted by such high profit of seigniorage, there will always be banks that cannot help issuing more notes than the gold reserves. Some may argue that big banks will value their credit and will not do so. However, the fact is that there have always been financial frauds. More importantly, financial fraud is just a non-systematic risk, which will only harm those who are deceived, while excessive issuance of bank notes will cause systemic risks such as inflation, which will harm all who are not deceived.

Moreover, money is special in that as long as one bank overprints notes, the whole economy will suffer inflation even if the other banks do not. In a modern society that has not been gold standard, it is difficult to judge if the quantity of notes printed by a bank is too much or too few. It is the total quantity notes printed by the whole banking system that can be judged to be too much or too few. Therefore, if any private banks can print paper money, it is not clear which bank or banks should be responsible for inflation. However, if the right of issuance is concentrated in one bank (central bank), it is clear that this bank is responsible for inflation. In other words, it is the best institutional arrangement for knowing the total quantity of money and who should be responsible with the least information cost.

Of course, government or central bank is also possible to be tempted by seigniorage to issue excessive paper money. It is precisely why inflation has become a common phenomenon in modern economy. As mentioned in previous section, even in the ancient times when it was very inconvenient to create inflation, there were emperors who had issued worthless base metal coins to make up for the fiscal deficit and eventually led to the collapse of the regimes. However, the transaction cost of monitoring tens of thousands of private banks not to issue excessive notes is undoubtedly higher than that of a central bank. There is no absolutely safe and error-free institution in the world, and human can only choose the one with the least transaction cost as far as possible. Obviously, it is the system that is most likely to stabilize the value of money that the central bank as the monetary authority to monopolize the issuance of notes. From the transition from the system that any private banks can issue paper money, which caused continuous economic crises to the system that central banks were set up by almost all the nations to monopolize the issuance right of money, it is sufficient to prove that the transaction cost of controlling the issuance by the market is relatively high, and no wonder human have chosen the current monetary system.

Section 7: The Truth of Bretton Woods System

In order to better overcome the difficulty of the issuer who is tempted by the high profits of seigniorage to overprint notes, the right to issue money is concentrated from numerous private banks to a central bank. However, private banks will overprint paper money, and so will a central bank. If a central bank issues notes strictly according to the gold reserves in the treasury, it is the true gold standard. Generally, there will be no inflation (unless there is a sudden great increase in the quantity of gold) and neither will be there economic crisis.

Since central banks in various nations have monopolized the power to issue notes, the super-inflation caused by excessive printing is not unique, and it has become common to print a little more to create small inflation. The reason to the former behavior similar to suicide is precisely the same as that of the trilogy of the collapse of the ancient Chinese dynasties. When a fiscal crisis is imminent, the central bank can only overprint money even it knows it is like drinking poison to quench thirst.

A typical example is the super-inflation in Germany after World War I and before World War II. After Germany's defeat, it was forced to pay huge war reparations to the victorious nations, which caused the Germany to fall into an extremely difficult fiscal situation. Foreign reparations must be paid in hard currency such as gold, but domestic fiscal expenditures can be paid by overprinting money. As a result, the same horrible phenomena as "gold yuan" also occurred in Germany, and the soaring prices had wiped out a large number of people's savings being overnight. In China, the super-inflation caused by "gold yuan" note led to the collapse of the regime of KMT party, while in Germany, it led to the rise and coming into power of Hitler and his Nazism.

On July 1, 1944, the Monetary and Financial Conference of United Nations that was later commonly known as the "Bretton Woods Conference" was held in Bretton Woods, United States. Although the Second World War was not completely over at that time, the Allies were sure that they would win the war. Therefore, their focus had shifted from the war to how to rebuild the new international order after the war. One of the important components of the new order was precisely the international monetary system.

All participants agreed that governments should not overprint money to fight against the fiscal crisis, and the international monetary system must return to the gold standard. However, most nations were faced with the problem that as two world wars broke out one after the other, the gold reserves originally held by rich European nations have almost all flowed into the United States. The US was not destroyed by the wars and had been continuously producing war materials for the other belligerent nations, so it had earned a large amount of gold as the international currency for settlement through international trade. As a result, most nations did not have enough gold reserves to issue money. The US proposed a compromise solution of the "Bretton Woods System" which meant the US that held a large amount of gold reserves issued US dollars that were directly linked to gold (at the ratio of 1 ounce of gold to 35 US dollars), and the currencies of other nations were linked to the US dollars, thus indirectly linked to gold. Under this international monetary system, the exchange rates between countries were still fixed, as in the gold standard.

However, there is some dirty trick in this Bretton Woods system! If most nations could not return to the gold standard because they did not have enough gold reserves, neither did they have enough US dollars at that time. It is precisely why the US implemented the European Renaissance Plan (commonly known as the "Marshall Plan" named by Marshall who was the marshal of US

army in World War II and became the Secretary of State after it). The essence of Marshall Plan was to give US dollars to European nations for free, but in the name of helping Europe to rebuild from the ruins of war. Objectively if the US had not given large amounts of dollars to European nations through the Marshall Plan, the Bretton Woods System would not work.

The question is: why does the US not directly give gold for free? If the US directly gave gold to European nations, they would have enough gold reserves to issue currencies and return to the gold standard. The reason is obvious: the US dollar can be printed at negligible cost, and of course it can be given for free. What is more, European nations would pay back the US dollars to buy goods and services from US firms for reconstruction. However, it is only a superficial reason. The US could also lend gold to European nations at low interest rates or even free of interest, and they could still pay the gold to buy goods and services from US firms in international trade, and then pay back the principal of gold gradually. Why not?

If the US truly wants to help European nations to rebuild their economies and the gold standard, the Bretton Woods system is actually unnecessary. What was the motivation of the US? Economics cannot speculate on such unobservable psychological activities. However, the objective effect shows that the US dollar have successfully become an international currency with the aid of the Bretton Woods system, taking the place of gold as an international currency for thousands of years. Although before the First Industrial Revolution and the World War II, Britain's economic strength had reached a peak and surpassed the rest nations of the world, and the British pound had accordingly become an international currency, it could not be a match for gold, let alone taking the place of gold. It is not natural that the US dollar has reached such important position as today.

At the end of World War II, there were little gold reserves for all nations. Now there are still little gold reserves for all nations, but they are holding a large amount of US dollars as foreign reserves. It is precisely why even if they are dissatisfied with the poor performance of the US dollar as an international currency, they could not get rid of it and return to the gold standard. In fact, the Bretton Woods system lasted only short time. Less than 30 years later, in 1971, the U.S. President Nixon's refusal to exchange US dollars for gold was the symbol of its collapse. However, the U.S. dollar as an international currency has remained unchanged.

Gold had been used as an international currency for thousands of years and doing a good job. It was the fault of governments of various nations that overprinted money and abandoned the true gold standard before World War II, not the fault of gold that it did not do a good job as an international currency. By contrast, the US dollar has been used as an international currency for only 60 years, but there has been an international currency crisis almost once in a decade, which shows that it has done an extremely poor job.

The reason is very simple. Firstly, the US dollar is paper money, so it is easily overprinted, and when slightly there will be inflation, and when seriously there will be economic crisis. Secondly, the US dollar is the national currency of the US, but it serves as the international currency of the world. As the central bank of the US, the Fed will, of course, preferentially consider the interest of the US. It is impossible to require it to implement a monetary policy that is good for the whole world at the expense of the US. For example, during the sub-prime mortgage crisis in 2008, the Fed continuously adopted the monetary policy of "Quantitative Easing" (QE) in order to stimulate the US economy. As a result, it had not caused inflation in the US, but caused global inflation in many other nations.

Although the Bretton Woods System claimed to be true gold standard, actually it had quickly turned into a pseudo one due to the massive issuance of US dollars as well as the "Triffin Dilemma". At the very beginning, the US dollars were rather scarce, so there was a "dollar shortage" in various nations. Then the Fed printed a large amount of money, and it quickly turned

into a "dollar glut" because the US dollars issued had been far more than 35 times of the gold reserves.

Just as the private banks overprinted notes out of line with the gold reserves in their warehouses at the very beginning, people would not find out because the disconnection was not serious. However, the disconnection would be getting worse and worse, so sooner or later people would find that the gold reserves held by the Fed could not support such a large amount of US dollars issued. Once they do not believe in the Fed any more, people would exchange US dollars for gold, resulting in a sharp drop in the gold reserves of Fed. Within the half month of March 1968, the gold reserves of the Fed flowed out of more than 1.4 billion US dollars. On March 14, in the gold market of London, the volumes reached a record of 350-400 tons. At last, in 1971, after the dollar crisis had repeatedly broken out 7 times, the Fed refused to exchange gold for US dollars, which was simply default! The Bretton Woods System actually collapsed.

There seemed to be a debt crisis of the US in 2011, and there was rumor that US bonds might lose the AAA rating. Some US officials placated the market and said that the US government had never defaulted on the debts. Did he forget, or did he not realize that 40 years ago the Fed refused to exchange gold for US dollars, which was precisely the biggest debt default in history! When foreigners accepted US dollars, they actually had established a contract that the US government promised them that the US dollar was a voucher of debts for the products or services they provided, and they could always get the money back at a rate of 1 ounce of gold to 35 US dollars from the Fed. However, suddenly the US government refused to honor this promise.

After the collapse of the Bretton Woods System, the currencies of various nations have completely lost the stability of the gold standard. Any government can print as much paper money as they like, and are only constrained by the conscience and discipline. As a result, domestically inflation has become a common phenomenon, while internationally the exchange system has changed from a fixed one to a floating one. The exchange rate risk has increased greatly, which undoubtedly affects the development of international trade.

After World War II, the developed countries, under Keynes's "instigation", justifiably used monetary policies to increase money supply (in fact, create inflation) to stimulate economic growth. Due to money illusion, the real economy can indeed grow temporarily when people do not realize there is inflation. However, sooner or later, people will know the truth. Even if government continues to create inflation, the economy still returns to the previous sluggish state, but now with inflation. In the 1970s, western countries that had been using Keynes's prescription for more than 20 years after the war, began to taste the bitter effects of it and had to face the dilemma of "stagflation" which meant there were economic stagnation and high inflation at the same time.

This phenomenon could not be explained by Keynes's theory, so monetary school that criticized Keynes attracted attention. In essence, the monetary school advocates to returning to classical economics that the value of money derives from goods and the increase in the quantity of money out of line with the increasing in the quantity of goods will only cause inflation. The goal of monetary policy should not be to stimulate economic growth, but to control inflation and create a stable environment for economic growth. Since then, in western countries this view of monetary school has generally been accepted.

However, because inflation does have the effect of stimulating economic growth temporarily, and the democratic election institution in western countries makes the politicians have the pressure to boost economic performance temporarily in election year, so central bankers cannot help but increasing money supply to create a moderate inflation under the pressure from politicians, which has been an open secret known to all. In fact, democracy encourages politicians to tend to take short-sighted actions, and developed countries are willing to create a moderate inflation even if there is no imminent fiscal crisis. It is precisely why a moderate inflation below 10% has

become such a common phenomenon in modern economy. It is also why Friedman as the leader of monetary school advocates the monetary policy of "single rule" that stipulates that the money supply should be increased at a long-term fixed growth rate (consistent with the long-term average growth rate of real economy). He had been trying to push for the "single rule" to be written into the constitution of US so as not to be influenced by politicians, and he failed.

Section 8: The Quantity Theory of Money

In this section, the second difficulty will be discussed. Even though the monetary authorities are willing to abide by discipline and do not overprint money, they actually do not know what the quantity of money is.

The example in section 3 is extremely simplified to illustrate the cause of inflation. If there are 100 units of goods, only 100 units of money should be issued. If the quantity of goods increases by 10% to 110 units, the quantity of money should also increase by 10% to 110 units. There will be no inflation. In reality, the problem is not that simple. Firstly, how to measure the quantity of goods? Different goods have different units of measurement, how to add up the quantity of goods with different units? Secondly, how to measure the quantity of money? It seems easier to measure the quantity of money than of goods, because money has a unified unit of measurement, but there is different difficulty in measuring the quantity of money.

From classical economics to neoclassical economics, it is quantity theory of money that has been prevailing. Fisher and Marshall had put forward "Fisher Equation"¹ and "Cambridge Equation" respectively, and they are logically similar, so we only introduce Fisher equation here: $PQ = MV$.

The left side of this equation reflects the monetary value of real economy and can be understood as the demand for money, because P is the price level (the average price per unit of good), and Q is the quantity of goods, which means there are so many goods in real economy that demand so much money. The right side of the equation reflects the quantity of money supplied. M is the quantity of money issued into the economy, and V is the times of turnover per unit of money (the velocity of circulation of money). Suppose one pays a bill of RMB100 in a transaction, it means this bill facilitates the transaction worth RMB100. The seller uses the bill to pay another transaction, which means the bill facilitates another transaction worth RMB100. Therefore, a bill can exchange for much more than the face value of it in a certain period of time, and should be multiplied by the times of changing hands.

Fisher equation is a tautology, so it must be true. The left side of the equation is actually the perspective of the buyer (how much has been spent to buy a good), while the right side of the equation is the perspective of the seller (how much has been received for selling a good), so it reflects the same transaction. If the equation is rewritten as $P = M \bullet \frac{V}{Q}$, under the constraint that the quantity of goods in real economy and the velocity of circulation of money are basically stable in short run, it is clear that the change in the quantity of money (M) will directly cause the change in the price level (P), which is precisely the mathematical expression on the cause of inflation.

However, if the quantity of goods in real economy and the velocity of circulation of money are not stable, what is the proper quantity of money in line with the quantity of goods? It is a difficult problem. For example, when there is super-inflation, people will exchange paper money for goods as soon as possible, so the velocity of circulation of money will be much higher than normal, which will intensify inflation.

1. Attention: it is different from the Fisher's Wealth Equation introduced in Lecture 18. [e](#)

Section 9: Money Multiplier

The greater difficulty lies in the fact that since there is banking system in modern economy, money multiplier makes the quantity of money more and more difficult to know.

Here is a numerical example to illustrate money multiplier. A central bank prints a note of RMB100. Suppose Person 1 gets it and deposits it in Bank A, so there is an increase of RMB100 as deposit in the accounting book of Bank A. A bank must lend out the deposits to make money, but it cannot lend out all the deposits, otherwise it will have no money to return the depositors who come to withdraw money. Every day, there are people who deposit money into the bank, while there are also people who withdraw or borrow money from the bank. Based on the experience of the bank's daily business, it can know how much money it should keep to meet the withdrawal of depositors every day.

Suppose there are 100 persons who deposit money into the bank and on average deposit RMB100 per capita, so the bank has RMB10,000 of deposits. Usually, these 100 depositors will not go to the bank to withdraw their money on the same day. Based on the experience, the bank knows that only one of them will withdraw money per day, which means it only needs to keep RMB100 to meet the withdrawal, and the rest of deposits can be lent out. The deposits kept in the bank to meet the withdrawal are called "Deposit Reserve", and the proportion of it to the total deposits is called "Deposit-Reserve Ratio (DRR)".

Suppose the DDR is 10%, so when Bank A has a deposit of RMB100, RMB10 will be kept as deposit reserve, and the rest of RMB90 will be able to be lent out. Suppose that RMB90 is lent to Person 2, and he deposits it in Bank B, so there is an increase of RMB90 as deposit in the accounting book in Bank B. Bank B keeps RMB9 as deposit reserve, and lends the rest of RMB81 to RMB81... This process can continue indefinitely.

Attention: bank deposit is also money! It should be easy to understand when thinking about the payment by a Debit Card. The money in a debit card is the money in the corresponding demand deposit account. Another example is the payment by check. As for the transactions between firms, most are settled through bank transfer. Therefore, not only the paper money that is in the hand is money, but also is the deposit in accounting book of bank in the form of numbers.

The money supply in a country is not only the paper money printed by the central bank, but also the deposits "created" by the commercial banks. In fact, the so-called money supply includes M0, M1 and M2 based on different statistical calibers. M0 is the money supply in narrow sense and M2 is that in broad sense. Paper money printed by the central bank is called monetary base or high-powered money, which is denoted by M0. The number in accounting book created by commercial banks through deposits and loans are called deposit money, which is denoted by M1 and M2. M1 is the sum of M0 and demand deposits, while M2 is the sum of M1 and time deposits.

Keynes believed that people like to hold cash because it has "Liquidity" which means cash is flexible as water and can become any other asset people need at any time. Of course, the disadvantage of holding cash is that it has no income like other assets (such as bank deposits, stocks, bonds, real estate, etc.), so cash is not profitable. Strictly speaking, other assets also have liquidity. For example, demand deposits can be withdrawn at any time, only a little more troublesome than cash because the depositors need to go to the banks. By contrast, time deposits cannot be withdrawn during the deposit term, or else the depositors cannot obtain the fixed interest originally agreed with the banks because they actually break the deposit contract and will be punished by only obtaining the much lower interest of demand deposit. Thus, the liquidity of it is worse than that of demand deposit.

As long as the stock market is not closed, stock-holders can sell their stocks at any time, so the flexibility in time is not worse than that of demand deposit. However, the price of stock will fluctuate, and the timing of selling out may coincide with the period when the price is falling. It is possible that the price can rebound after a while, and then it will be more favorable to sell out. But the stock-holder is eager to have cash and cannot wait, so he has to sell out the stock even when the market condition is unfavorable. He is sure to suffer a loss if the price of stock is lower than that he bought in. Even if it is higher, there may be opportunity cost as long as the price is rising, so the present price is still lower than the expected future price. By contrast, bank depositors will not suffer any loss in the principal, and will lose the interest at most if the time deposits are withdrawn in advance. However, a stock may be sold when the price is lower than that when it was bought, which means the stock-holders will suffer a loss in the principal. It is why the liquidity of stock is worse than that of bank deposit.

The liquidity of real estate is even worse because it takes quite a long time to be sold out, while there are active transactions of stocks on every working day. If an estate-owner cannot wait for a long time, he has to offer a much lower price to attract buyers as soon as possible, so he will suffer a much greater loss than the stock holder or the bank depositor.

As cash has the best liquidity, the liquidity of other assets is determined according to the convenience of cash realization (assets are turned into cash) and the degree of loss the holder will suffer in the principal and interest. The more easily an asset can be turned into cash, and the less loss the holder will suffer in the principal and interest, the better the liquidity. In fact, the lower the transaction cost of an asset that is turned into cash, the better the liquidity.

The difference between M1 and M2 precisely lies in the fact that the liquidity of demand deposit is almost the same as that of cash (paper money). Although the liquidity of time deposit is a little worse, it can be withdrawn in advance, and the loss is only limited to the interest, which means it can easily become demand deposit. It is why both M1 and M2 are regarded as money.

Because the liquidity of money is much higher than other assets, the term "liquidity" is usually used as another name for money. For example, the sentences such as "the central bank releases/returns liquidity" often appear in financial reports, which means the central bank increases/decreases the money supply. Strictly speaking, liquidity is only the nature of money, not money itself, and all assets have a certain degree of liquidity, which is not unique to money.

Back to the above numerical example, with the first deposit, the deposit money created by Bank A is RMB100. With the second deposit, the deposit money created by Bank B is RMB90. With the third deposit, the deposit money created by Bank C is RMB81... and so on. Arranging these numbers together, there is an infinite diminishing geometric progression: 100, 90, 81 ..., and each value is 90% of the previous value. According to the summation formula $\frac{a_1}{1-q}$, the total amount of deposit money created by the banking system is $\frac{100}{1-90\%}$ or 1000. To generalize this formula, it is $\frac{M_0}{R}$ (R is the DRR). $\frac{1}{R}$ is the money multiplier because it reflects how many times of deposit money is created by commercial banks from the monetary base issued by the central bank.

Thus, the money supply in a country is almost the deposit money created by commercial banks, and the paper money printed by the central bank accounts for only a small part, even a fraction, if R is very small and the money multiplier is very large. Of course, the central bank can still directly or indirectly control the money supply by directly controlling the monetary base and affecting the money multiplier through changing R . In all countries, there are regulations on "legal deposit reserve ratio", so commercial banks cannot keep as many deposits as reserves as they will.

Section 10: The Monetary Policy Tools of Central Bank

In economics textbooks (including the textbooks such as "Money and Banking"), three major monetary policy tools of central bank are listed: the adjustment of the legal DRR, the adjustment of the rediscount rate, and the Open Market Operation.

(1) The adjustment of the legal DRR

The adjustment of the legal DRR refers to adjusting R, thus changing the money multiplier. However, in all economics textbooks, it will be supplemented quickly that this monetary policy tool is a blockbuster that is rarely used. In fact, the central banks initially implemented this legal DRR not as a monetary policy tool, but as a means of controlling the risks of commercial banks. Reserve deposits are those that banks cannot lend out, so there is an opportunity cost for banks and they are reluctant to keep too many reserve deposits. If there are not enough reserve deposits, there will be a risk that depositors cannot withdraw their. In addition to implementing monetary policies, the functions of a central bank also include supervision of commercial banks, so it implemented a mandatory legal DRR to prevent the banks from keeping too few reserve deposits.

Later, when economists studied monetary theory, they found that the DRR is an important factor that directly determines the money multiplier, so the adjustment of the legal DRR can be a monetary policy tool in theory. In fact, central banks rarely use this tool, but the central bank of China has frequently used it after 2008, which seriously harmed Chinese economy. The purpose of monetary policy is to stabilize the value of money and provide a stable environment for economic growth. Even if DRR is only slightly adjusted, it will greatly change the money supply through the amplification effect of the money multiplier, which causes economic fluctuation.

What is more, the adjustment of legal DRR also seriously interferes with the business decisions of commercial banks. DRR should be determined by the proportion of deposits withdrawn from the banks every day, so the unnecessarily high legal DRR will cause a large amount of deposits to be idle and the banks to be unable to make normal profits, which will reduce the efficiency and is actually an increase in transaction cost.

(2) The adjustment of the rediscount rate

The term of "discount rate" has been introduced in Lecture 18, but there it is used to convert the future value into present value, while here it refers to the interest rate charged by a bank when a firm sells commercial paper to it.

A firm can provide short-term financing to a buyer in the form of "accounts receivable" in transactions, which means a buyer can receive the goods without paying immediately until three months, six months or other different periods later. In essence, the firm has lent a sum of money equivalent to the payment of the goods to the buyer, and the loan period is three months, six months, etc. The buyer can also issue a commercial bill to the firm, indicating the amount and the date of the payment on it, and then the firm's accounting entry is not "accounts receivable", but "bills receivable". The firm can choose to hold the bill until the maturity date to request the buyer to honor the bill, or sell the bill to a bank before the maturity date, which means it has transferred the bill to the bank, and the bank will request the buyer to honor the bill on the maturity date.

Since the bank has bought this commercial paper, it needs to pay the firm. In theory, the payment should be based on the face value (the payment of the goods), but the bank needs to make money on this operation, so it will make a discount on the face value before making payment. The difference between the money actually paid by the bank to the firm and the face value of the bill is actually interest, because the bank allows the firm to get the payment before the maturity date, which is equivalent to lending the money to it, and it is the buyer who will repay the bank on the maturity date. The interest rate in this case is also called "discount rate" that is not essentially different from the term introduced in Lecture 18 because it does not represent the discount rate of future value (the payment on the maturity date).

Since the discount rate is the interest rate at which a firm borrows money from a bank, the rediscount rate is the interest rate at which a bank borrows money from the central bank. If a firm is short of money, it borrows from a bank. What if a bank is short of money? As mentioned in Lecture 18, there is an interbank market, in which a bank can borrow money from other banks. It can also borrow money from the central bank, because the central bank can print money and it is sure to have money. In fact, the central bank is called the "lender of last resort" of the whole economy, because it is responsible to provide the monetary base for an economy.

The term "rediscount" refers to that a bank sells bills to the central bank. It cannot sell commercial bills, because the central bank is a government agency rather than a commercial institution and cannot participate in the operation of firms. The bills sold by a bank are central bank bills or treasury bonds that are securities issued by the central bank or the Ministry of Finance. Since the discount rate is the interest rate at which a bank lends money to firms, the rediscount rate is, of course, the interest rate at which the central bank lends money to banks.

If the central bank raises the rediscount rate, it means the interest rate of loans paid by banks to the central bank rises, so they will also raise the interest rate of loans to firms. Rising interest rate will inhibit firms from borrowing money, thus reducing investment and production, at last affecting economic growth, so it is a tight monetary policy. On the contrary, if the central bank cuts the rediscount rate, the market interest rate will also decline, which will encourage firms to borrow money, thus increasing investment and expanding production, at last stimulating economic growth, so it is an expansionary monetary policy.

However, the adjustment of rediscount rate is also rarely used in western countries now, because the central bank is in a passive position and has to wait for commercial banks to borrow money. Especially as the interbank market is booming, commercial banks often borrow money from peers unless it is tight money in the whole banking system and the interest rate in the interbank market is also very high, or there is a problem for the bank, so the credit risk of default has increased greatly, and no one is willing to lend money to it in the interbank market.

(3) Open market operation.

It refers to the operation of the central bank to release or recover money by buying and selling securities (mainly foreign exchange, treasury bonds, central bank bills, etc.) in the open market (including foreign exchange market, monetary market, etc.). If the central bank wants to implement an expansionary monetary policy which means releasing money into the economy, it will buy securities in the open market, and release money by paying money. On the contrary, if the central bank wants to implement a tight monetary policy which means recovering money from the economy, it will sell securities in the open market, and recover money by collecting money.

Compared with the above two monetary policy tools, open market operation has the advantages of activeness, flexibility and timeliness. The central bank can fully control the scale of operation and has considerable activeness. Open market operation is flexible, because the central bank can adjust the direction of monetary policy and the money supply at will. The timeliness is also very

good, because the central bank can implement the monetary policy immediately and continuously as long as the market is open.

The monetary policies of various nations are mainly implemented through open market operation. For example, 90% of the money release or recovery of the Fed is carried out through open market operation.

Section 11: The Difficulty in the Target of Monetary Policy

Open market operation is only a monetary policy tool, not the monetary policy target. As far as the ultimate target of monetary policy is concerned, all the central banks want to control the price level (inflation), but the question what should be the intermediate target of regulation is quite controversial.

Monetary school has always advocated money supply as the intermediate target, but since Greenspan became the chairman of the Fed, he liked to regulate the interest rate in the federal fund market (the interbank market of the US). The selection of intermediate target should consider measurability, controllability and correlation. The interest rate in the interbank market is so obvious that it does not have to be measurable, and it is also controllable through open market operation. However, correlation is the biggest problem.

Those who take interest rate as the intermediate target believe that the rise in interest rate reflects the decline in money supply relative to demand, and vice versa, so the level of interest rate is related to the quantity of money supply. However, the interest rate that changes along with the change of money supply cannot be the pure interest rate. A grain sowed into the soil can produce more grains a year later. How can the growth rate of this grain be affected by the increase or decrease of money supply? As mentioned in Lecture 18, the pure interest rate excluding inflation and risk (transaction cost) is the rate of return on investment that is objectively determined by the production capacity or efficiency in an economy, and cannot be changed by the monetary policy. As for the interest rate in reality, it does fluctuate under the influence of the money supply and demand, which actually reflects the change in the understanding of people about the pure interest rate and the factors of inflation and risk. The Fed interferes with the interest rate in the federal funds market by open market operation, which in fact cannot change the pure interest rate, but only makes the interest rate in the market no longer reflect the pure interest rate correctly, thus actually interfering with the understanding about the pure interest rate, which implies a rise in the information cost.

Although the interest rate in reality does have some relationship with the money supply, the regulation on it will interfere with the understanding about the pure interest rate, which is the same mistake as creating inflation to interfere with the understanding of people about the real price, and will also cause economic fluctuation or business cycle. The most typical example is the sub-prime mortgage crisis in the US in 2008. Greenspan frequently changed the interest rate during his reign, making it sometimes rise and sometimes fall just like waves, causing economic entities sometimes to expand and sometimes to shrink their consumption and investment (production), which had nothing to do with the changes in real economy. Although the frequent fluctuation in interest rate was not the main cause of the financial crisis in 2008, it was sure to be a big "accomplice".

How about directly regulating the money supply? The second difficulty is precisely that the monetary authority often does not know what the quantity of money (money supply) is! According to Fisher equation, the effective money supply is to multiply the quantity of paper money by the velocity of circulation. According to money multiplier, paper money is only M₀, and most of money supply comes from deposit money created by commercial banks. However, $\frac{1}{R}$ is only a theoretical maximum of money multiplier instead of an effective value in reality.

Firstly, because the legal deposit reserves must be turned over to the central bank, commercial banks have to keep a certain amount of deposit reserves called "excess reserves" to meet the withdrawal of depositors. The amount of excess reserves varies for different banks, but it is sure to increase the effective DRR, leading to a decrease in money multiplier, and also a decrease in the quantity of deposit money.

Secondly, from the numerical example of money multiplier in Section 9, it is obvious that in the process of creation of deposit money by commercial banks' operations of deposit and loan, depositors cannot withdraw cash. If a depositor withdraws part of cash, the money will leave the banking system and not continue to participate in the creation of deposit money, which also leads to a decrease in the quantity of deposit money. It is called "leakage" that refers to the withdrawal of cash that causes money to leak out of the banking system, resulting in a decline in the ability of commercial banks to create deposit money.

Thirdly, commercial banks can only create deposit money through the continuous operations of deposit and loan. If the operations are generally interrupted, the money multiplier will become invalid, resulting in a sudden sharp drop in deposit money.

Why do banks not engage in loan business? Generally, banks are sure to lend out the deposits to make money, but in some extreme cases, banks will be reluctant to lend. The most typical extreme case is when the economy is in recession, people are very pessimistic about the future. Consumers dare not borrow to spend, so consumer credit declines. Firms are unwilling to invest in production, so do not borrow money from banks. Banks neither believe that the investment of firms can be successful, so even if they apply for loans, banks are afraid that the loans will become bad loans and refuse to lend. It did happen in the financial crisis of the US in 2008. In order to stimulate economic growth with an expansionary monetary policy, the Fed had cut the federal fund interest rate to 0, but the loan rate of commercial banks was still above 5%. Obviously, most of the rate is risk premium, reflecting a rise in the risk of business failure, so the (loan) interest rate could not fall to 0.

There is a concept of "Liquidity Trap" invented by Keynes, which means that no matter how low the interest rate is during an economic depression, people will not increase the demand for money, leading to the failure of monetary policy. However, his explanation is wrong. The correct explanation should be as followed. Firstly, poor economic performance implies higher risk, and the part of risk premium in the interest rate will rise, so it is impossible for the loan rate to drop to 0. Secondly, the increase in risk also leads to banks' reluctance to lend, which will interrupt their normal loan business and cause money multiplier to fail, so the quantity of deposit money decreases sharply. Even if the central bank injects a large amount of monetary base into the economy, it accounts for only a very small proportion of the money supply, so several times or even more of increase in monetary base may not make up for the decrease in deposit money and the total money supply has not increased. Thirdly, even if the total money supply has increased, because people are extremely pessimistic about the future and dare not consume or invest, the more money just lies in warehouses and does not circulate in the economy¹, so the effective money supply has actually decreased.

As an international currency, the US dollar brings about a bigger problem that the money released by the Fed can flow out of the US, which is leakage in broad sense. The withdrawal of cash causes money to leak out of the banking system, which is leakage in narrow sense. If there is leakage in broad sense, the US dollars flowing out of the US cannot be effective money supply. During the financial crisis in 2008, people had no confidence in the economic prospect of the US, but had greater confidence in other countries (such as the emerging markets), so they invested the US dollars released by the Fed in the markets of other countries. As a result, even if the Fed had carried out QE many times, the inflation rate in the US did not, but it had caused serious inflation in other countries.

US dollars that flew into other countries would cause the central banks of other countries to passively issue more money. Take China as an example, foreigners who want to invest in China should first exchange US dollars for RMB in the foreign exchange market, which means they sell out US dollars and buy in RMB. It will lead to an oversupply of US dollars, so the exchange rate has to change to eliminate the disequilibrium supply and demand. It is, of course, the depreciation of US dollars and the appreciation of RMB. In order to maintain the stability of the exchange rate of RMB, China's central bank will sell out RMB in the foreign exchange market to absorb excessive US dollars, thus passively issuing more RMB, which is called "Position for Foreign Exchange Purchase". Foreigners use the RMB exchanged from the foreign exchange market in consumption and investment in China's domestic market, causing an increase in the quantity of RMB or inflation in China.

The global inflation during the financial crisis in 2008 was mainly caused by the QE of the Fed. The purpose of the Fed is to create domestic inflation, but the pessimistic expectation of the US economy caused the more issuance of US dollars by the Fed not to stay in the US to effectively increase the domestic money supply, but leak out of the US to other countries, thus increasing the money supply of other countries. There had happened political instability and regime change in many Arab countries (the so-called "Arab Spring"), the source of which actually the QE of the US. Take Egypt as an example. Why did the Egyptians overthrow Mubarak? Do they really want democracy and anti-dictatorship? Don't be naive! Mubarak had ruled Egypt for more than 30 years. If Egyptians really hated his autocratic rule, should not they have overthrown him long ago? The truth is: there was serious inflation in Egypt during that period. There had been extreme disparity between the rich and the poor in the Middle East countries. The governments relied on generous welfare to appease the poor. However, the QE of the Fed caused serious inflation (with the inflation rate of over 20%) suddenly, and welfare payments were often fixed and would not immediately be adjusted along with inflation. In fact, people with fixed incomes are those who suffer the most from inflation. It is precisely why the poor could not survive anymore and rise up.

The inflation in China also became serious during that period, but was much lighter than other countries, which was attributed to the strong growth of China's real economy. Although the money supply had increased passively, the real economy had also increased, and some of the increased money was absorbed by the growth of the real economy, so as to keep the inflation rate below 10%.

In a word, although the central bank can directly control the monetary base and indirectly control the deposit money created by commercial banks through the legal DRR, the effective quantity of money in the economy is also affected by many factors beyond the control of it. There is still the second difficulty: how to know the effective quantity of money (money supply) in a country? The monetary base issued by the central bank is only a small part of money supply, and the main part of it is deposit money created by commercial banks that is also affected by many factors. If it is an international currency (like US dollar), there is even the possibility of leaking out of the country.

From the next section on, the best monetary system that can solve the two major difficulties of monetary policy will be introduced.

1. In 2016, there was "negative interest rate" in Japan, which meant those saved money in banks could not have positive interest but had to pay interest to banks. The Japanese government wanted to "force" people not to save money, but to spend money on consumption or investment. However, after the implementation of this policy, the sales volume of household safes had increased greatly, which implied that people only transferred the money from the warehouses of banks to safes at home, and still refused to spend, so the money leaving the banking system had not circulated in the economy. ↩

Section 12: The Formation of Zhu's Monetary System

The best monetary system comes from the improvement of the monetary system formed during the period of Zhu Rongji in charge of the central bank in China. To understand Zhu's monetary system, we must first know how it came into being.

After Deng Xiaoping's southern tour in 1992, China had been in another period of rapid economic growth. However, during this period, there had been serious inflation until 1996 due to the prevalence of power lending. The CPI(Consumer's Price Index, an important index commonly used to measure inflation rate) rose as high as 14.7% in 1993, and in 1994, it even reached to the highest 24.1% since 1949.

Why was the inflation rate so high in China at that time? According to the monetary theory introduced in this lecture, the direct reason must be that too much money had been issued and was out of line with the growth of real economy. But why? Why did the Chinese government not control the growth of money supply? Because the powerful men and their relatives and friends (the so-called "Prince's Party") relied on their power to borrow money from banks. At that time, although the People's Bank of China had already performed the functions of a central bank (including printing paper money, controlling money supply, and supervising financial institutions) as state agency instead of a commercial firm, it still engaged in the business of commercial banks including loaning. Thus, when the powerful men and their relatives and friends wanted to do business, they would force the People's Bank of China to lend money to them.

A central bank is a state agency, and should engage in commercial loan, but the central bank system of China was not perfect at that time. Here, by the way, we will introduce the history of the development of banks in China after 1949. During 1949 and 1978, there was only one bank (the people's bank of China, PBC) in China besides the bank of China that specialized in foreign business. PBC was both a state agency and a commercial bank, which was a typical example of mixing up the functions of the government and enterprises. After the reform and opening up in 1978, PBC had gradually separated some of the operations of commercial banks out to form three of the four major state-owned banks in China (Industrial and Commercial Bank of China, Agricultural Bank of China and China Construction Bank). However, until 1993, PBC was still able to lend money out, so the powerful men and their relatives and friends to put pressure on it to lend money to them. As a result, PBC had involuntarily lent out a lot of loans, which increased the money supply and created serious inflation at that time.

Because of the bitter experience of overprinting "gold yuan" note on the eve of the collapse of the KMT regime before 1949, the Chinese people abhor super-inflation and the Chinese government is also deeply aware that the consequence of super-inflation will affect social stability and even overthrow the national regime. Zhu was appointed as the president of PBC in June 1993. Shortly after he took office, Zhu vigorously promoted the reform of PBC by completely stopping the commercial banking business of loaning. On March 18, 1995, the Law of PBC was passed. For the first time, the status of PBC as the central bank of China was clarified in the form of national legislation. Since 1996, the inflation rate had dropped sharply. In 1998 and 1999, there was even slight deflation according to CPI. In fact, there was more serious deflation than what CPI reflected.

At the same time, the exchange rate between US dollar and RMB dropped from more than 5.5 in 1992 to about 8.6 in 1994 when inflation was at the peak, implying China's high domestic inflation rate. However, after that, although the inflation rate dropped sharply and even turned to deflation, the exchange rate did not fluctuate significantly and had remained stable at 8.27 for a

long time since 1997. As all know, 1997 was the year when the Southeast Asian financial crisis broke out, all currencies in Southeast Asian countries depreciated sharply against the US dollar, including the Thai baht depreciating 39%, Korean won 36%, Indonesian rupiah 72% and Singapore dollar 61% respectively. RMB of China not only remained stable during the crisis, but even appreciated slightly compared with 8.29 in 1996.

Why could RMB withstand the impact of Southeast Asian financial crisis? Although Zhu was no longer the president of PBC in 1997, he had set the mechanism of stabilizing the exchange rate between the US dollar and RMB at about 8 before he left. A considerable part of China's economic growth comes from export, and the main trade partner of China is the US, so the stable exchange rate between RMB and US dollar is important to maintain a stable foreign exchange environment for export enterprises in China. As mentioned in Section 3, internal inflation is precisely the same as external currency depreciation, so Zhu stabilized the exchange rate between the US dollar and RMB when he tried to control domestic inflation in the mid 1990s by adopting a system of pegging to the US dollar.

One may ask, during the financial crisis in 1997, the Southeast Asian countries also wanted to stabilize the exchange rates between the US dollar and their currencies, but why could China make it while they could not? The secret is the Southeast Asian countries used the foreign exchange reserves of US dollar to stabilize the exchange rates, while China used monetary policy!

Suppose there are international speculators attacking a country's currency, which means they sell out the country's currency in large quantities in the foreign exchange market, causing the supply to exceed the demand and the price (the exchange rate between the US dollar and the currency) to fall. The central bank usually fights against the attack of the speculators by selling out the foreign exchange reserves of US dollar to buy in the domestic currency, equalizing the supply and demand to keep the exchange rate stable. However, a country's foreign exchange reserves are limited and will run out quickly, and then the central bank can do nothing but surrender to speculators and let the currency depreciate. By contrast, PBC or China's central bank does differently. It uses the monetary policy of reducing the money supply domestically! Precisely because it is a domestic monetary policy, there is no need to use foreign exchange reserves of US dollar.

Zhu came up with this system to stabilize the exchange rate between the US dollar and RMB not to fight against the speculators during the financial crisis in 1997, but to control domestic inflation in 1993 when he became the president of PBC. Because what he was thinking about was how to decrease the inflation rate, the tight money policy to reduce the money supply was, of course, the direct solution. Since internal inflation is the same as the external currency depreciation, it is natural to invent this system of directly controlling inflation and indirectly stabilizing the exchange rate by monetary policy.

When the Southeast Asian financial crisis in 1997 broke out, the foreign exchange reserves in China were as high as 139.89 billion US dollars, almost double of those two years ago (1995). Thailand, as the first Southeast Asian country to be attacked by the speculators, had only 00-35 billion US dollars as foreign exchange reserves. In the first half of 1997, 20 billion US dollars were sold out in the foreign exchange market to stabilize the exchange rate between the US dollar and Thai baht, and more than 4 billion US dollars were sold out in May alone. Chinese government had witnessed how the speculators attacked the currencies of Southeast Asian countries and knew that even such foreign exchange reserves might not be sufficient to resist the speculators.

In fact, even if the speculators were not taken into account, it was estimated that Indonesia, Malaysia, South Korea, Thailand and the Philippines had experienced a sharp change from a net inflow of 93.8 billion US dollars in 1996 to a net outflow of 24.6 billion US dollars in 1998, reversing the capital by more than 100 billion US dollars. In other words, just the flight of the

private capital (in order to avoid the risk of depreciation, rather than deliberately attacking the currencies) was enough to exhaust almost all the foreign exchange reserves in these countries.

Therefore, the Chinese government wisely chose not to sell out a large amount of US dollar in the foreign exchange market to fight against the speculators, but to firmly control the money supply domestically by tight money policy, keeping the exchange rate between the US dollar and RMB at the level of 8.27. RMB is the currency of China, so PBC has full autonomy in releasing and recovering it, which is different from that the other Southeast Asian countries used their reserves of US dollar to fight against the speculators. The US dollar is the currency of the US, and cannot be as many as these Southeast Asian countries wanted to have.

Thus, China not only stabilized the exchange rate between the US dollar and RMB, but also did not cost a cent of foreign exchange reserves of US dollar in 1997, with the system invented by Zhu.

Section 13: From the Monetary Policy Regulating Economy to the Monetary System Stabilizing Money Value

According to the statement that internal inflation is the same as external currency depreciation, why shouldn't Zhu's system of stabilizing the exchange rate with monetary policy be used to control inflation? In fact, it is precisely the argument of monetary school that monetary policy should not be used to stimulate economic growth, but only to control inflation or keep the value of money, which means monetary policy has become monetary system.

However, is it not what most developed countries do now? The central bank keeps an eye on CPI to adjust monetary policy, which means it will adopt tight money policy when there is high inflation rate, and easy monetary policy when low inflation rate. What is special about the monetary system of Zhu? Zhu's system will be introduced in detail.

Zhu's system of stabilizing the exchange rate with monetary policy is right in general direction, but it is necessary to be improved on an important detail. The general direction refers to pegging a clearly visible target (anchor) by increasing or decreasing the money supply to stabilize the value of money. The central bank as the issuer of money is completely independent in increasing or decreasing the money supply, and does not need to be subject to anything uncontrollable by it (such as the reserves of US dollar). Therefore, this "pegging" is feasible.

But an important detail is what the target is. As mentioned in Section 11, the intermediate target must be "measurable, controllable and correlative". With the open market operation as the monetary policy tool, there is no problem for the requirement of controllability, so there are the requirements of measurability and correlation left. For measurability, the target must be clearly visible. If it were not, how to peg to it? It is like when one is shooting, he needs to aim at the target, but if the target is not clearly visible, how can he judge whether he has aimed and hit it? Obviously, Zhu chose the exchange rate between the US dollar and RMB that is a clearly visible target because there is foreign exchange market to make all can see the exchange rate every minute and even every second, and it is accurate to four digits after decimal point.

Besides, the target must reliably represent the value of money for the requirement of correlation, because the value of money derives from that of goods it exchanges for. For a good target, when money has pegged to it, the value of money has also pegged to the value of goods. Zhu's system of pegging RMB to US dollar cannot meet this requirement. The US dollar is only another currency, not goods. Unless the value of US dollar is stable, which means the exchange rate between the US dollar and goods is stable, and a direct peg to the US dollar is equivalent to an indirect peg to goods, Zhu's system can really stabilize the value of RMB instead of just the exchange rate between the US dollar and RMB.

It was lucky that during Zhu was the president of PBC and then the Prime Minister of China, the value of US dollar had been rather stable. As known to all, Clinton was the president of the US from 1993 to 2000, and during his term, the economy of US had experienced a long period of high growth and low inflation that was called "new economy". Besides, during much the same period (1991-2005), Greenspan was the chairman of the Fed, and had been adopting the "strong dollar" policy of controlling inflation domestically and keeping the US dollar strong in the foreign exchange market. Therefore, the value of the US dollar value during this period had been stable and strong, which made it a very good target for pegging.

However, good time did not last long. Since 2003, as the economy worsened in the US, while that in China continued to grow, the exchange rate between these two currencies needed to represent the competitiveness of countries behind, so the U.S. dollar was under pressure to depreciate against RMB. However, as RMB was pegged to the US dollar by Zhu's system, the US dollar could not depreciate against the RMB to reflect the fact that the economic strength of the US had declined relative to that of China. From then on, the US had been criticizing China to manipulate the exchange rate, requiring it to increase the value of RMB. The underlying reason for it is precisely that the US dollar has no longer been stable, and not a good target for RMB to peg.

In fact, it cannot be a long-term solution to peg to another country's currency, unless it also pegs to a good target, which means pegging to it is equivalent to pegging to the good target. After all, the value of money derives from that of goods, not from that of another country's currency. Thus, what is a good target?

From ancient times until the US dollar has become an international currency after the World War II, the general monetary system was gold standard. If it is the true gold standard system, gold is precisely what money is pegged to. For example, if it is stipulated that the central bank can only issue RMB100 when there is 1 once of gold in the warehouse, the price of gold (the exchange rate between gold and money) will be stabilize at the RMB100/ounce. It can be regarded that RMB has pegged to gold.

As long as the central bank issues money strictly in accordance with the above ratio, no speculators can change the price of gold, and neither can they change the value of RMB. Suppose there are 1 million ounces of gold in the warehouse, the central bank will issue 100 million RMB. There is only so much RMB in the market, and the speculators can only get up to 100 million RMB to exchange for gold. When they exchange all the 1 million ounces of gold from the central bank, which means the central bank has no more reserves of gold, but neither will there be any RMB in the market, so the speculators will not be able to sell out more RMB in the market to change the exchange rate.

In the financial crisis of 1997, the Southeast Asian countries ran out of foreign exchange reserves under the attack of speculators and had to surrender to depreciate their currencies, because they did not strictly issue money according to the exchange rate between their currencies and the US dollar, or else it was impossible that there were not enough reserves of US dollar. In other words, Southeast Asian countries had adopted a "pseudo dollar standard".

By contrast, besides the RMB of China had been successfully defended, neither had Hong Kong's dollar depreciated, not because China declared that if Hong Kong's foreign exchange reserves were insufficient, China would lend the foreign exchange reserves to support Hong Kong, but because Hong Kong's monetary system was "Linked Exchange Rate System". It refers to that Hong Kong Monetary Authority issues Hong Kong dollar at a ratio of 1:7.8 (US dollar to Hong Kong dollar) strictly based on the reserves of US dollar. In other words, Hong Kong has adopted a "real dollar standard", which is logically similar to the real gold standard, except that the standard currency is US dollar instead of gold, so Hong Kong dollar is pegged to US dollar.

However, although both RMB and Hong Kong dollar are pegged to US dollar, Zhu's system is different from Hong Kong's monetary system, because PBC does not issue RMB based on the reserves of US dollar. There must be reserves of US dollar in Hong Kong's linked exchange rate system, but it is not necessary for Zhu's system. Take the gold standard as an example, and suppose the change rate between gold and RMB is 1:100. If it were Hong Kong's system, the central bank could only issue RMB100 when it holds 1 ounce of gold. However, for Zhu's system, the central bank does not need to hold any gold, but keeps an eye on the price of gold in the market. If it rises to be higher than RMB100/ounce, the central bank should adopt tight money policy to decrease the quantity of RMB in the market relative to gold, thus causing the price of gold to fall. If the price of gold falls to less than RMB100/ounce, the central bank should adopt

expansionary monetary policy to increase the quantity of RMB in the market relative to gold, thus causing the price of gold to rise.

There are several advantages for Zhu's system compared with Hong Kong's system.

(1) The most direct advantage is that the central bank does not need to hold reserves (gold, US dollar, or anything as the target of money).

(2) The money supply in an economy is actually far more than the paper money (monetary base) issued by the central bank, and most of money is the deposit money created by commercial banks. Although the quantity of deposit money is related to that of monetary base, it is not directly controlled by the central bank. Therefore, it is possible that the central bank wants to adjust the money supply by changing the quantity of monetary base, but that of deposit money does not change as expected, so neither does the money supply change as expected, which make the price level still unstable. In addition, for the international currency such as US dollar that can "leak" out of the country, it is even more difficult for the central bank to effectively control the money supply.

In other words, if the central bank issues monetary base based on the reserves, because the intermediate link from monetary base to deposit money is extremely complicated, and for the international currency that can "leak" out of the country, it is very difficult to stabilize the price level by controlling the effective money supply. Therefore, it is better not to calculate what the appropriate money supply is, but to directly stabilize the price level. As the above example of gold standard, with Zhu's system, the price of gold is directly stabilized by monetary policy instead of issuing money based on the reserves. The essence of Hong Kong's linked exchange rate system is to peg the quantity of money to that of reserves, while the essence of Zhu's system is to peg the value of money to the price of target.

(3) The value of money derives from the goods it exchanges for, and what money exchanges for is more than gold. Gold, as a good rather than a currency, is useless, so pegging to gold is actually similar to pegging to another currency such as US dollar. Just the US dollar is a kind of paper money that can be overprinted, while gold is a precious metal that cannot be as much as one's will, so it is not easy to create inflation under gold standard. It is precisely why gold had done well as an international currency for thousands of years, while the US dollar has been overprinted in only half a century. As money, gold is inherently stable, while all paper money is easily overprinted to create inflation.

Gold as money is still not a good, and the value of money does not derive from it but goods. There are so many goods in the world, and some of them are perishable, how can they be directly used as the reserves. What is more, when multiple goods are used as the targets of pegging money, what should be the exchange rate between money and the multiple goods? Obviously, Hong Kong's system is simply not feasible if multiple goods are directly used as the targets, but Zhu's system is.

Section 14: The Monetary System

Pegging Money to the Price Index of a Basket of Tradable Goods

The central bank can select some goods that are important to life and production, puts them into a basket in a certain proportion, and calculates the price of the basket of goods. For example, RMB100 can buy the combination of goods in this basket. The central bank does not need to keep the basket of goods as reserves, but only needs to keep an eye on the price of it. If the price rises, the central bank should adopt tight money policy to make the price fall. If the price falls, the central bank should adopt expansionary monetary policy to make the price rise. Take the gold standard as an example. It is actually a special case of the basket of goods, which means there is only one good (gold) in that basket, so the price of the basket is that of gold.

Some may wonder if it is price control. No! What is controlled here is the price level, not the relative price (real price) of goods. There is a numerical example to illustrate that this monetary system pegging to a basket of goods has not influence the relative prices of goods.

Suppose the initial price of pork is RMB10/kg and that of rice is RMB5/kg, and there are 5 kg of meat and 10 kg of rice in a basket with the price of RMB100. If the price of pork rises to RMB20/kg, the price of the basket will also rise to RMB150. The central bank should reduce money supply to make the price of the basket fall from RMB150 back to RMB100. Assume that the new price of pork after monetary policy is x and that of rice is y , the equations should be $\frac{150}{100} = \frac{20}{x}$ and $\frac{150}{100} = \frac{5}{y}$, and the solutions are $x = \frac{40}{3} \approx 13.33$ and $y = \frac{10}{3} \approx 3.33$. Due to the change of money supply, the price of the basket returns to RMB100, while the price of pork correspondingly falls from RMB20 to RMB13.33, and that of rice from RMB5 to RMB3.33.

Obviously, the central bank does not control the price of meat and rice. What it does is to control the value of money. Although the monetary prices (nominal prices) of the goods in the basket rise or fall to keep the total price be RMB100, the relative price of them has not changed. Before the monetary policy adjusting the money supply, the price of the basket is RMB150, and the relative price of pork and rice is 4: 1 (20: 5). After the monetary policy making the price of the basket fall back to RMB100, the relative price of pork and rice is still 4:1($\frac{40}{3} : \frac{10}{3}$). Therefore, the monetary system pegging to a basket of goods is not price control at all, and what the central bank changes is only the monetary price of good.

Fixing the nominal price of a good through monetary policy is totally different from the price control mentioned in Lecture 9. The following means seem to fix the price of a good: (1) Administrative instruction directly prohibiting suppliers from pricing above the regulated price (price control), (2) Demand control (such as rationing system and purchase restriction policy), (3) Increasing supply, and (4) controlling money supply.

No economist has ever regarded the means (4) as price control, because the essence of it is to stabilize the value of money, and what is controlled is actually the price of money, not that of any good. The price of any good should not be controlled except that of money, because one of the functions of money is a ruler of value and the length of a ruler should not be changing. In other words, those who doubt that the monetary system pegging to a basket of goods is price control are wrong in criticizing the means (4) with the consequences of means (1).¹

One may notice that the price of this basket of goods is actually a price index calculated by multiplying the price of the goods in the basket by their proportions. The CPI is also a price index based on a basket of goods (consumption goods). What is the difference between them?

Let us introduce the calculation of CPI and how to use it to measure inflation first, and then explain the defect of it. Continue the above example of the basket of pork and rice, and suppose it is precisely the basket that makes up CPI. Take 2010 as the base period when the price of pork is RMB10/kg and that of rice is RMB5/kg. If the price of pork rises to RMB20/kg in 2012, CPI is the current price of a basket of goods divided by that in base period: $\frac{150}{100} \times 100 = 150$. If the CPI of 2011 is 140 that is must also be based on 2010, the inflation rate from 2011 to 2012 is $\frac{150-140}{140} \times 100\% = 7.14\%$.

In reality, the basket of goods that makes up CPI cannot be so simple. Take China as an example, food accounts for 31.79%, alcohol and tobacco 3.49%, housing 17.22%, communications 9.95%, personal medical and health products 9.64%, clothing 8.52%, household equipment and maintenance services 5.64%, entertainment, education and cultural products and services 13.75% of the CPI basket. As for the base period, it changes every five years, keeping the same cycle as China's "five-year plan". For example, the CPI from 2011 onwards is based on 2010.

The monthly CPI data in financial reports reflect the year-on-year growth, which are calculated by comparing with the CPI in the same month of last year, thus directly equivalent to inflation rate. For example, the CPI in March 2012 rose 3.6% on year-on-year base, which is calculated by comparing with the CPI in March 2011. The other data is on month-on-month base, which is calculated by comparing with the CPI of the previous month (February). The data on year-on-year base rather than on month-on-month base can eliminate the influence of seasonal factors. For example, in January or February, there is Spring Festival in China with much more consumption than the other months in a year, which will push up the price level, resulting in a lower month-on-month growth in CPI of the following month.

In addition to CPI, there is also a price index called PPI (Producer's Price Index) that is calculated by the factory price of the product (instead of the retail price directly faced by consumers). Obviously, PPI is closely related to CPI and is theoretically a leading indicator of CPI.

Many countries use CPI to measure inflation rate, so to control inflation with monetary policy is to control CPI. In other words, the currencies of these countries are theoretically pegged to the price of a basket of goods that makes up CPI. So what is so special about the revised version of Zhu's system that the target is changed from the exchange rate to the price index of a basket of goods?

The key lies in the fact that the target must be clearly visible (measurable). There are many problems in CPI.

(1) The basket that makes up CPI contains goods with varying quality. As known to all, the prices of such goods as pork with different qualities can be rather different. The case of real estate is even more serious. A same house, built in prime locations of a city or in remote villages, will have totally different price. The absolute difference of quality is not so important, because CPI measures the change of price during a period of time, so long as the goods keep the same quality during that period, there will be no problem. However, the problem is precisely that the quality of goods will also change during the period. With economic development and technological progress, the quality of goods will usually rise, and the price should also rise, not because there is more money supply. If a society's production and technological capabilities have generally improved, the qualities of the goods will generally increase, so the prices will generally rise. CPI will mistakenly measure it as inflation. For example, from 1996 to 2000, the CPI in China showed a slight deflation in China (such as -1.4 in 1999). In fact, the qualities of goods in China generally increased during that period, so the data of CPI seriously underestimated the deflation.

(2) It is difficult to deal with the new products for CPI. If there are new products such as computers and mobile phones, money can buy more kinds of goods, which implies the purchasing power or value of money has increased. Therefore, CPI will overestimate the inflation because it again mistakenly measures the rise in value of money as inflation.

Attention: a rise in price is not the same as inflation. Inflation is caused by the increase in the quantity of money out of line with that in the quantity of goods, and it will but it will lead to a rise in price. However, a rise in price is not necessarily due to inflation, but may also be due to the general rise in quality or the appearance of new products. If all rises in price are regarded as inflation, to control inflation may mistakenly become suppress quality improvement and product innovation. In addition, at the beginning of the reform and opening up, China had experienced a general rise in price during the period of abolishing the price control. Neither was it was inflation, but the controlled price returned (increased) to the market equilibrium.

(3) The statistics of CPI is very time-consuming, so the CPI of a month cannot be known and released to the public until the following month, thus causing a time lag of recognition in monetary policy.

In other words, CPI is not clearly visible enough to be a measurable target. What the central bank needs is a good price index like the exchange rate in the foreign exchange market or interest rate in the interbank market that all can see immediately every minute and even every second. The goods put into the basket must be carefully selected. One of the principles for choosing the goods in the basket is that the quality of the goods is stable and clear, so that the prices are visible every minute and even every second. The other principle is that the goods are representative, which means they play an important role in a country's life and production, so they can represent the value of a large number of goods.

Those most appropriate are commodities in futures market and some in wholesale market (such as cement). Although there are hundreds of millions of commodities in the world, the raw materials needed to manufacture them are basically those in futures market and wholesale market (such as oil, steel, cotton, etc). As long as the prices of these raw materials are stable, those of hundreds of millions of commodities are also stable. Moreover, even if there are new products, the raw materials used to manufacture them are not new, so the problem of new products CPI cannot deal with is not a problem anymore. The central bank does not need to collect the information of prices of millions of commodities, but only needs to collect that of hundreds of raw materials that is visible in futures market and wholesale market every minute and even every second!

Thus, the basket of goods is selected, and the proportions of these raw materials are determined according to their proportion in a country's production and life. The central bank then calculates the price index of the basket, and keeps an eye on it to adjust the money supply to stabilize it on the desirable level. The central bank can use computers to connect with major futures markets and wholesale markets in the world, directly input relevant price data from those markets, and publicize the price index of the basket every minute, so that the public can see clearly that RMB is firmly pegged to this index. It is as feasible as the index of foreign exchange market, and the fact that China successfully fought against the international speculators in the financial crisis of 1997 has eloquently proved the effectiveness of this monetary system.

Once RMB is pegged to the price index of the basket of tradable goods, the exchange rate control can be completely abolished, allowing the exchange rate between RMB and all foreign currencies to float freely. If the exchange rate changes, it reflects the change in the value of foreign currency instead of RMB, because the value of RMB has been tightly pegged to the value of the basket and will not change.

It is precisely the best way for the internationalization of RMB. After pegged to the price index of a basket of tradable goods, foreign exchange control can be abolished, so RMB becomes a

convertible international currency. The control on the exchange rate between RMB and the US dollar can naturally be abolished. If there is a depreciation of US dollar, it is the problem of US dollar, not the appreciation of RMB. And the US has no reason to accuse China of manipulating the exchange rate any more. For the other countries, the best choice for them must be to peg their currencies to RMB instead of the US dollar, because the value of RMB is pegged to the real source of money. As long as China strictly abide by the system of pegging to the price index of a basket of tradable goods that is publicized every minute, there will never be any risk of overprinting RMB, so there will never be depreciation of RMB.

Reviewing the two major difficulties of a good monetary system, we can better understand the advantages of this monetary system pegging money to the price index of a basket of tradable goods.

For the first difficulty, if a government is determined not to abide by the discipline of stabilizing the monetary value, there is no way for any monetary system to stop it. However, as long as a government pretends to abide the monetary discipline, under this monetary system, it will be very difficult for the government to overprint money secretly. Under the fiat monetary system, people do not know whether the government has secretly overprinted money until the inaccurate CPI is released at least a month later. By contrast, under this monetary system, the price index of the basket is visible every minute and even every second, which makes it very difficult for the government to cheat.

For the second difficulty, this monetary system cleverly goes round the difficulty of knowing the quantity of money directly. It monitors the price index of a selected basket of goods with stable quality instead of knowing the specific quantity of money. A water supply company also faces similar difficulty in natural science. It has to transport water to tens of thousands of households through water pipes, and the water pipes from the water supply company to the households are also very complicated. It is impossible to accurately know the specific quantity of water needed to ensure that each household has water flowing out when turning on the tap. The water supply company does not try to know the specific quantity of water needed, but widely installs water pressure gauges, and then continuously supplies water into the pipe until the water pressure meets the standard. If the gauges indicate that the water pressure is lower than the standard, the water supply company will inject more water into the pipe to increase the water pressure. If the gauges indicate that the water pressure is higher than the standard, the water supply company will draw water back from the pipe to reduce the water pressure. Thus, even if the city continues to expand and the pipe network has to expand with it, the water supply company does not need to continuously get to know the specific quantity of water needed, but only needs to keep an eye on the water pressure gauges to ensure that the water pressure is stabilized at the standard level. It is logically similar to the monetary system pegging money to the price index of a basket of tradable goods which converts information difficult to know to that easy to know, thus greatly reducing the information cost of stabilizing the monetary value.

So far, it is the monetary system with the least comprehensive transaction cost, so is the best one!

1. The price control criticized by economists do not include means (4), neither include means (3). In fact, increasing supply is the most correct means to fix price that is effective and has no adverse consequences. However, increasing supply is usually subject to resource constraints, production cycle and etc., so it is often impossible to be adopted. [2](#)

Lecture 20: Income Distribution, Wage Contract, Unemployment

Section 1: Income Distribution: Marginal Productivity Theory

Those who know the textbooks of MSE will notice that after the topics of producer theory and market structure, this book does not turn to introduce the factor market or turns from the resource allocation to income distribution, as the ordinary textbooks will do.

The reason has been explained before that there is no essential difference between factors (raw materials or intermediate products) and final products (consumption goods). Final products can be regarded as goods for consumption, and can also be regarded as raw materials or intermediate products used to produce labor as factors. In this lecture, it will be further explained that as long as the transaction cost is low enough and piece rate contracts are used, product market is inseparable from factor market. Thus, the analysis on product market can be analogized to factor market, so there is no need to elaborate on it, and to put forward some key points.

(1) The equilibrium of $MR = MC$ is universally applicable, not only to different market structures such as price-taking and price-searching, but also to factor market. When this equilibrium is applied to factor market, the specific form can be written as $mrp = h$, where mrp refers to marginal revenue of product that means the increase in income brought about by an additional unit of production factor. In other words, mrp is an alternative form of MR , calculated by mp (marginal product) multiplied by P (price of product). As mentioned in Lecture 15, the curve of mrp is the demand curve of production factor, so the shape sloping downward is determined by the law of diminishing marginal product. And h is the price of renting production factor, so it is an alternative form of MC . The economic implication of this equilibrium is that the income (rent) obtained by each production factor is equal to the marginal contribution to production of it.

Why should the contribution be observed at margin? Because all factors have contributed to production, but how to separate the contribution of one factor? If only one factor is input to production without any other factors, it is even impossible to carry out production. However, if there are other factors input, the contribution of this one factor will be mixed with that of other factors. The solution is "Ceteris Paribus" that means keeping the input of other factors unchanged and only changing the input of this one factor, and then the increase in income (the mrp of this one factor) must be attributed to the contribution of this one factor (because other factors have not increased and cannot be the source of income increase).

There are some assumptions for the equilibrium of $mrp = h$. Firstly, there is no monopoly in the product market and the price of product remains unchanged. Secondly, there is no monopoly in the factor market and the price (rent) of factor remains unchanged. Thirdly, the quantities of other factors are assumed to remain unchanged.

If the first two assumptions are relaxed, the analysis will become complicated, but only the specific form of $MR = MC$ is changed, which will not affect this equilibrium. For example, Fig.20-1 reflects the case when there is monopsony in the factor market. If compared with the Fig.16-1 (that reflects the case when there is monopoly in the product market) in Lecture 16, it is obvious that they are logically similar. The curve of MC in Fig.16-1 corresponds to that of mrp in Fig.20-1, while the curve of MR corresponds to that of mfc , and the curve of AR (that is the same curve as D , the demand curve) corresponds to that of afc . The point H point in Fig.16-1 determines the equilibrium output (Q_m), which corresponds to the point of M point in Fig.20-1 that determines the equilibrium quantity of factor (a^*). The point of F in Fig.16-1 determines the equilibrium price of product (P_m), which corresponds to the point of N in Fig.20-1 that determines the equilibrium price of factor (h^*).

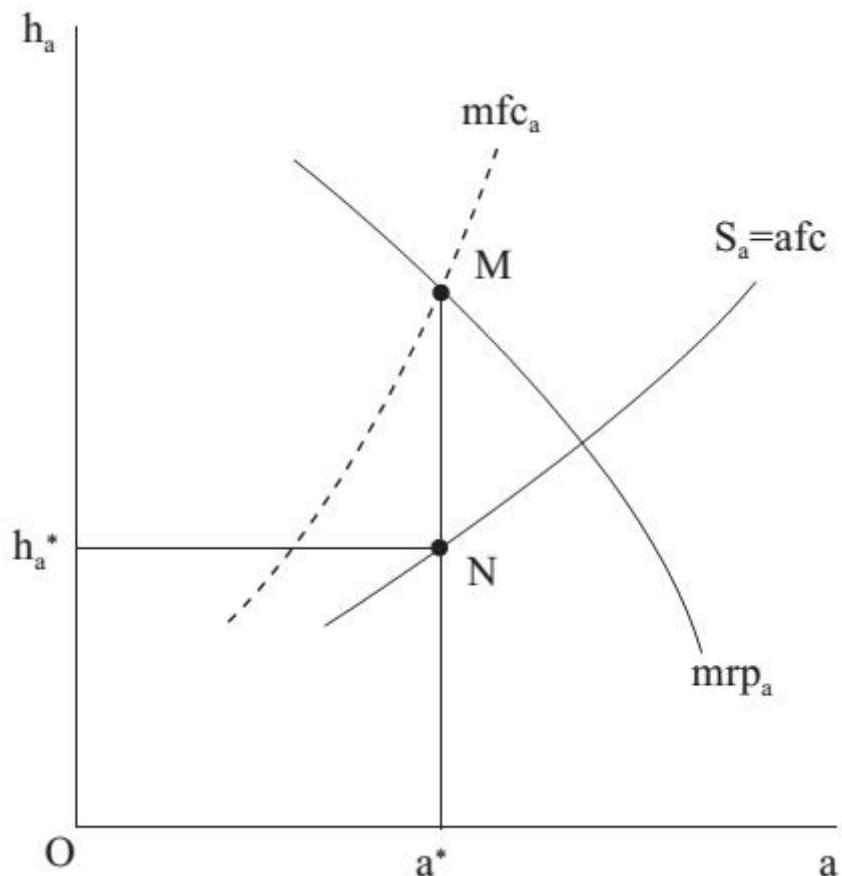


Figure 20-1

Besides, if the price of product price and that of factors change, it is not necessarily because there is monopoly in these two markets, but it is also possible that there is external diseconomy mentioned in Lecture 15, which still only changes the specific form of $MR = MC$.

If the third assumption is relaxed, the analysis will also become complicated, but what is important is that it cannot refute the law of demand for production factor, because it is guaranteed by the iron law of the diminishing marginal product¹, unlike the law of demand for product as an axiom established by postulation.

(2) With the Euler Theorem in Mathematics, it can be proved that all the income obtained from the sale of product is divided up by all the factors in production, which means there is no "surplus value" as Marx said, and it is divided according to the proportion of the contributions of the factors.

Neoclassical economics usually divides production factors into capital and labor. Thus, there is $Pq = mrp_L \bullet L + mrp_K \bullet K = wL + rK$, where w is the price of labor (wage rate), and r is the price of capital (interest rate). It is precisely the "Marginal Productivity Theory" which is the theory of income distribution advocated by neoclassical economics. Classical economics pays more attention to agricultural production, so there is also the factor of land, and the income of it is rent. However, as has pointed out in Lecture 12 and 18, if the concepts of rent and capital are generalized, the income of all factors is rent, and any good that can bring income is an asset, the market value of which is capital. In fact, labor is often also called human capital.

As shown in Fig.20-2, the curve of mrp (the demand curve of production factor) is sloping downwards, and the point where the price of factor (h) is equal to it determines the equilibrium quantity of factor (a^*). The price of factor price multiplied by the quantity of it is the income of it (the rectangular $OhEa^*$). NhE , the triangle (or curved triangle) formed by the curve of mrp above the horizontal line of price and the vertical axis is the consumer surplus mentioned in Lecture 15. Because it is the factor market, the so-called consumer is the producer who rents production

factors, and the so-called consumer surplus is actually the rent obtained by the other production factors. The sum of NEa^*O and NhE is the total income (gross output or rent) of the product.

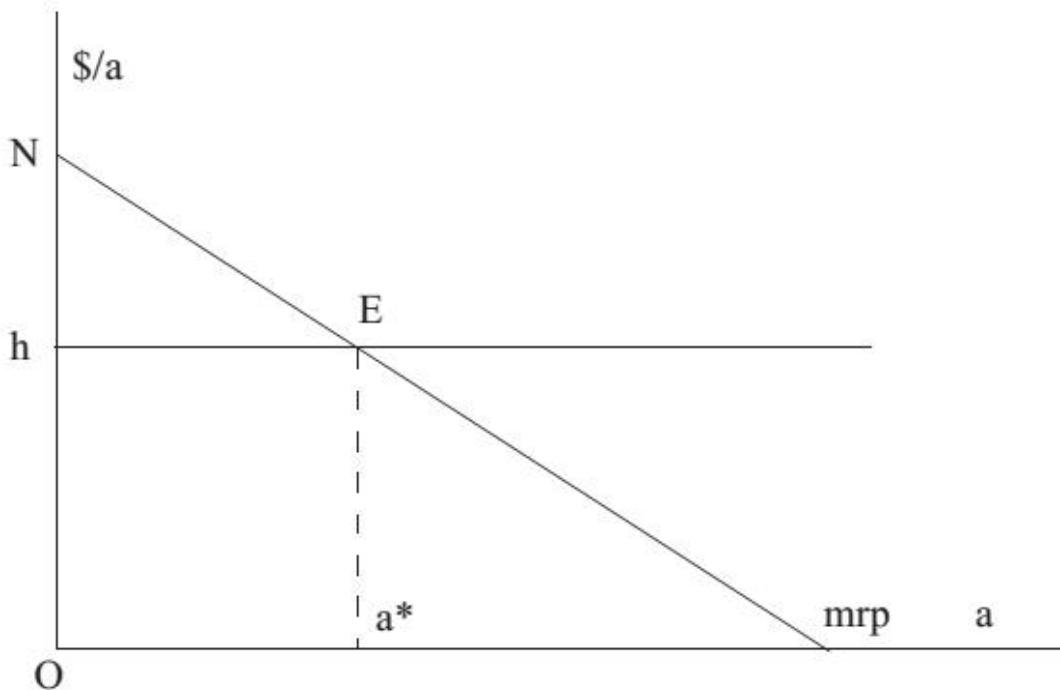


Figure 20-2

(3) When the textbooks of MSE introduce the supply curve of labor, there is often a graph in Fig.20-3 that with the increase of income, labor will increase the supply (such as working hours) at the beginning, but later will reduce the supply, resulting in the supply curve of labor bending backward. However, the supply curve is actually the same line as the demand curve, so the supply curve of labor bending backward has violated the law of demand that is axiomatic and does not allow any violation.

There is actually a very simple mistake in the analysis of MSE. The wage cannot be regarded as the net income of the labor gets before it is subtracted by the cost of becoming tired, getting ill due to the too hard work. In other words, as the wage rises, a labor's supply will increase, but his fatigue will also increase. After a certain point, the benefits of wage rise will be offset, so that the net income turns to fall, and the supply of labor will correspondingly decrease. It is logically similar to a progressive ad valorem tax levied by the government on a commodity. The net income of a seller is not the tax-inclusive price paid by consumers, but the difference of the price and the tax. As the price rises, the tax that the seller has to pay per unit of the commodity will also increase. After a certain point, the net income (after-tax price) that the seller gets turns to fall, and the seller will correspondingly decrease the supply.

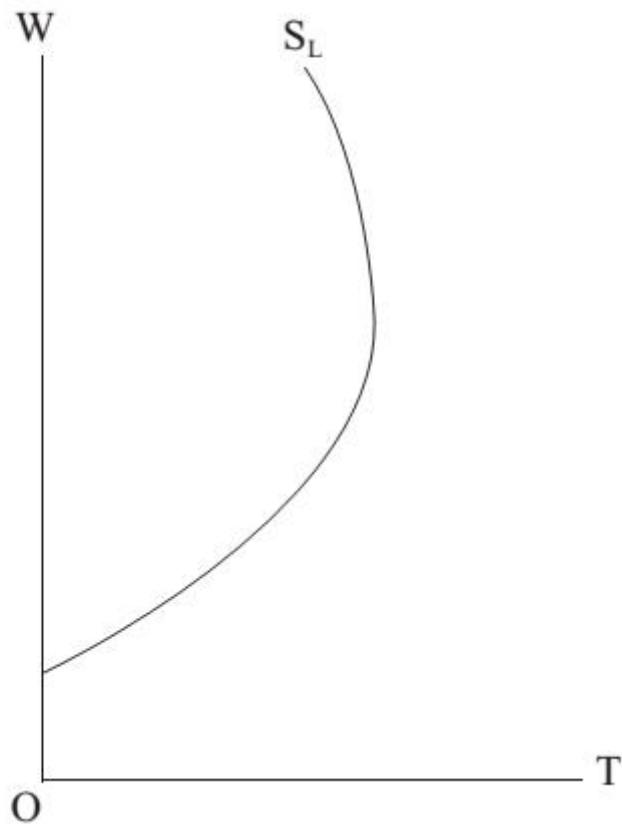


Figure 20-3

In order to eliminate the effect of offset by rising fatigue on the rise in wage, it is the entire labor market instead of single labor that should be observed. The rise in wage will attract more people to enter this occupation. No matter how high the wage rises, it will not cause a fall in the number of employment of this occupation (unless the rise in wage is caused by the offensive or risky factors that are similar to the rise in fatigue offsetting the benefits of rising wage). Therefore, the supply curve has not bent backward, and has neither violated the law of demand.

1. Even if all factors increase in proportion, marginal product at most keeps unchanged, and will not increase. [e](#)

Section 2: Wage Contract

The analysis in Section 1 is on the general factor market. In this section, the analysis will focus on the labor market, a specific factor market, which is also applicable to other factors, and the characteristics of labor are most obvious.

Why is labor singled out? As mentioned in Lecture 14, although in neoclassical economics production factors has been divided into capital and labor, the major difference between labor and non-labor is actually not understood. Labor is human with self-will, so there will be different transaction costs if it is organized in production with different contracts. The analysis on the general factor market in Section 1 is the traditional analysis of neoclassical economics without taking transaction cost into account, and it is only correct when the transaction cost is low enough, so the phenomena that can be explained are extremely limited. As for the phenomena that there are different contractual arrangements in seemingly similar production, it is entirely incapable.

One of the most important reasons is whether the transaction cost is low will determine whether the factor market can be separated from the product market. Labor market is a very good example to illustrate it.

Suppose one sits by the street as a shoebblack, is the price paid by a client used to buy the output of shining shoes (product) or the input of the labor (production factor)? The answer is: either! A shoebblack is self-employed and not a firm. In such a case, the product market is obviously inseparable from the factor market.

At the very beginning of Lecture 11, as has been criticized, the traditional producer theory (Theory of Firm) is confused with the Firm Theory in MSE. The traditional producer theory does not actually take the influence of the organizational form of a firm on production into account because there is no the concept of transaction cost in it. In other words, the producer theory only studies the case of individuals self-employed, where the product market is inseparable from the factor market.

In reality, why are there many productions organized through firms instead of market? The reason has been explained in Lecture 14 that under certain constraints the transaction cost of using firms to organize production is less than that of market. In other words, as long as the transaction cost of market is low enough, people will not choose to organize production through firms, so individuals self-employed instead will organize production through market, and the factor market is not separable from the product market.

One may ask: is it because the producers directly face the consumers that the factor market is inseparable from the product market? The answer is NO! The above example of a shoebblack does be the case that a self-employed producer directly faces consumers, but even if it is not the case, it does not mean that the factor market will become separable from the product market. There is another example of "stringing beads": the self-employed producers stringing beads do not directly face consumers, but sell the beads directly to them through middlemen (agents). In such a case, there are middlemen between producer and consumer, but the factor market is still inseparable from the product market, because the middlemen are in the same position as the consumers of the previous example of a shoebblack. Is the price paid by a middleman used to buy the output of beads (product) or the input of the labor (production factor)? The answer is still: either! The middleman just passes the beads on to the final consumers, and the consumers pay for either the output of beads (product) or input of the labor of those stringing the beads and the middleman investigating the market demand, conveying the information of this demand, transporting the beads to the market, and pricing (production factor). It is logically similar to the wholesale market.

Obviously, wholesale market is not factor market, and retail market is not product market, although the buyers in the former are usually middlemen.

The key is that the contract forms are basically the same. When a middleman buys the beads from the self-employed individuals stringing beads, the price is set on a piece-by-piece basis. Similarly, when a final consumer buys the beads from a middleman, the price is also set on a piece-by-piece basis. In other words, the price (wage) of labor, a production factor, is added into the beads on a piece-by-piece basis, which is called "piece wage".

One may ask: is it because the producer is self-employed that the factor market is inseparable from the product market? The answer is still NO! Just imagine there is a significant technological progress happened in the bead stringing industry, a large machine has been invented to greatly increase the production efficiency. The machine is too large to be set up in a family, so the self-employed individuals stringing beads sit around the machine in a big house to cooperate.

The middleman still buys beads from them on a piece-by-piece basis, but the price may include the rent of the machine and the big house. In fact, when the self-employed individuals work separately in their own homes, the price of the beads includes not only the price of labor price (wage), but also the cost of small tools and the rent of the workplace at home, and just the PPRs of these other factors clearly belong to them. In other words, the self-employed individuals rent out not only their own labors, but also small tools and the working place of their homes.

As for the large machine and big house, to whom the PPRs of them belong depends on the specific constraints. If the self-employed individuals rent the large machine and the big house, the price of beads will include the rent of both, but should be distributed to them according to the proportion of their capital contribution. If they have different production efficiency and capital contribution, this calculation will be very troublesome (implying there is high transaction cost). As a result, in reality, usually it is the middleman who rents the large machine and the big house, so the price he pays for the beads does not include the rent of both. Of course, when he sells the beads to final consumers, he will raise the price to include not only the wage of his own labor, but also the rent of both. In order to facilitate the self-employed individuals to cooperate, the middleman will probably set up a firm to employ them as workers, but the contract for buying the beads from them is still on piece-by-piece basis. Thus, from the perspective of law, there is a firm, but from the perspective of economics, it is not essential different from the case that the individuals stringing beads are self-employed.

Let us review the example of secretarial service in Lecture 14. The services of a secretary are various and trivial. Although every service can be priced on a piece-by-piece basis in principle, the transaction cost is obviously too high. It is more convenient for the middleman to simply price according to the working hours instead of the quantity of products, which means the wage contract has changed from piece rate time wage. Thus, a final consumer buys the product from the middleman on a piece-by-piece basis, while the middleman directly buys the working hours of the labor, which implies that now the product market is separated from the factor market. From the perspective of economic, there is a firm, and the middleman also has turned to the manager.¹

In a word, the key lies not in whether the producer directly faces the consumers, nor in whether the producer is self-employed, but in the form of wage contract. If the contract is on a piece-by-piece basis (such as the wage contract of piece rate), the product market is inseparable from the factor market, regardless whether the price only includes the wage of labor or also includes the rents of other production factors. However, if the contract is not on a piece-by-piece basis (such as the time wage), the product market is separated from the factor market, because in the product market it is the contract on a piece-by-piece basis that set a price for the product, while in the factor market, it is the contract of time wage that set a price (wage) for the labor. As mentioned in Lecture 14, Coase advocates that it is the organization of firm replaces the market

system to reduce transaction cost of market, which is actually not accurate enough. The more accurate statement is that it is the factor market that replaces the product market, and the most accurate statement is that it is the contract of production factor replaces that of product.

Think deeply, what a manager really wants to buy is always the products, not the working hours of the workers. Just because the transaction cost of directly pricing the product is too high, he turns to price the product indirectly according to the time. Suppose one can output 10 units of products in an hour, and the price of each unit is RMB1. The piece rate is RMB1 per unit of product, which is the same as that the time wage is RMB10 per hour. The essence of the time wage is to entrust the price of product to the time, because there are many kinds of products or services (such as the service of secretary) with too high transaction cost for pricing on a piece-by-piece basis. If the information that the price of each unit of product is RMB1 and there are 10 units produced in one hour is difficult to know, a manager will turn to use the time wage of 10RMB per hour. However, entrusting the price of product to the time can reduce the transaction cost of pricing on a piece-by-piece basis, but it will also increase another kind of transaction cost that workers will shirk during working hours, resulting in the less output in an hour than 10 units under the wage contract of piece rate. The loss caused by this decrease in efficiency is precisely an increase in transaction cost.

Of course, the manager will take some measures to control the workers. For example, he can require a worker to produce at least 10 units in an hour, or else the time wage will be cut. However, it means the manager still needs to know the information that a worker can produce 10 units in an hour, which is precisely the transaction cost caused by the wage contract of piece rate in such occupation as secretary.

Besides, there is still the problem of quality. A worker may produce 10 units in an hour, but the quality is poor, so he does not work as hard as he does in producing higher quality goods. The loss caused by the decline in quality is also an increase in transaction cost. The manager then has to specify the quality of the product in detail, which also needs to be monitored under piece wage. Under piece wage, only the quantity of products has been specified and monitored, so the workers will tend to increase quantity at the expense of reducing quality. The manager can supervise the workers himself or send someone to do so, but the supervisors need to be paid, which is also an increase in transaction cost.

In conclusion, different forms of contracts (direct pricing or indirect pricing by entrusted to some easy to measure such as time) can reduce some kinds of transaction cost, but will also cause the some other kinds of transaction cost to increase. Self-interested human naturally choose the contract with more transaction cost reduced and less transaction cost increased. In addition to the contracts of piece wage and time wage analyzed above, there are also sharing contracts (such as commission, dividend and bonus), and these different contracts may be used in combination (for example, the contract of time wage and sharing contract are often used in combination).

For any contract, one or several "property" will be chosen as the measure index for direct pricing. For example, in the contract of piece wage, the quantity of products is measured to set price, while in the contract of time wage, the working time is the measure index. The measure cost of these indexes chosen for direct pricing (transaction cost for pricing) must be low. Because they are directly and easily measured, the transaction cost for supervising their performance is also relatively low. However, for those properties that have not been directly measured and priced (such as the quality of products under piece wage and the shirking behaviors of the workers under time wage), the transaction cost for supervising their performance is relatively high, sometimes is so high that the supervision has to be given up, and the transaction cost is the loss in income caused by non-performance. In order to reduce the transaction cost of supervision, properties of direct measurement can be increased, but the transaction cost of measurement will also rise. Self-interested human will choose to increase the properties of direct measurement

until the increase in transaction cost of measurement is equal to the decrease in transaction cost of supervision on the margin.

To sum up, when the transaction cost of using the market is low enough, the product market is separable from the factor market. The price paid by the final consumers is either for the product or for the production factor. In other words, the price of the production factor is actually determined by that of product. As an important deduction, the wage of a worker is also determined by the income that the goods he produces can get in the market, not by the manager or boss of the firm. If government intervenes in the labor market, such as implementing the minimum wage law or giving trade unions administrative monopoly power, it actually indirectly intervenes in the product market.

When the transaction cost of using the market rises (such as the measurement cost of direct pricing the product), people will turn to direct pricing by other easily measurable properties of production factor to replace direct pricing, thus separating the factor market from the product market. However, the separation of the two markets does not mean that there is no linkage between the price of product and that of production factor any longer, but just becomes complicated. The price of the production factor (such as the wage of labor) ultimately still depends on the price of the product, so the government interventions in the factor market will eventually interfere in the product market.

The properties of direct measurement and pricing are stipulated by the price clause or income clause in a contract, which is similar to the transaction in the market, and guided by the invisible hand of the market without need of supervision. By contrast, the properties if indirect measurement and pricing are stipulated by the use clause in a contract, which is not guided by the market, so need to be guided and supervised by the visible hand of the manager, thus causing an increase in the internal management cost (the transaction cost of using the firm).

In a word, contract is structural that there are not only price clause (income terms) but also use clause that is used to supervise the prosperities of indirect measurement and pricing.

1. This imaginary example of significant technological progress happened in the bead stringing industry has a "prototype" in reality -- the industrial revolution. The essence of the industrial revolution is not only the major technological progresses of the inventions of large machines such as textile machines and steam engines, but also the emergence of factory system that replaced the traditional family workshop. [\[2\]](#)

Section 3: The Theory of Rent-encroaching

The analysis of general factor market in Section 1 is classified as microeconomic in the textbooks of MSE. The analysis of labor market in Section 2 belongs to new institutional economics. In this section, we will continue to analyze the labor market, but what will be analyzed is unemployment that is classified as macroeconomics in the textbooks of MSE. However, as mentioned in Lecture 6, it is not necessary to divide economics into micro and macro.

Let us discuss a problem between new institutional economics and the so-called micro and macro-economics first: if a government policy that intervenes in the market causes the price of a production factor to be higher than the equilibrium determined by the market, it will violate the marginal productivity theory (the theory of income distribution), because the income obtained from that production factor is higher than the contribution of it, which can only be achieved by encroaching on the income of the other production factors. Therefore, the essence of the government policy that intervenes in the market is to infringe on PPRs (the right to income) of other factors, which is precisely "rent-encroaching". Rent is the income obtained from an asset (production factor). Some factors obtain higher income than their contributions, which must be at the expense of encroaching on the rent of the other factors. For example, when government has implemented the minimum wage law, the workers with the equilibrium wage lower than the minimum wage will encroach on the rent of the other factors.

However, rent-encroaching is best observed from the perspective of the whole society rather than a specific firm. A firm can choose to dismiss the workers who are not worth getting the high salary to avoid losses, so rent-encroaching is best observed from the decline in output and wealth caused by unemployment in the whole society. If government further implements unemployment benefits, the government expenditures are in fact ultimately burdened by the taxpayers.

Therefore, although firms dismiss the workers and avoid rent-encroaching by not paying them wages, the taxpayers of the whole society have to pay them unemployment benefits, and their rents are encroached on by the unemployed. Unemployment benefit is usually lower than the minimum wage level, so it seems that the rent-encroaching on the taxpayers is lower than that on the firms. However, if the firms continue to employ the workers, they have made some contributions even if they are lower than the minimum wages. By contrast, the unemployed get welfare benefits without working, and they make no contribution to the society. Therefore, it is uncertain which loss in rent is higher.

How about a society without unemployment benefits? Regardless of the grievances of the unemployed that may lead to social instability and increase the governance cost (also a kind of transaction cost in broad sense), there are two options for the unemployed: to simply rest at home, or to peddle on the streets.

If the unemployed choose to rest at home, it is essentially different from the voluntary unemployment. For the voluntary unemployed, the value of leisure is higher than what they can earn from work. The rent of the whole society will not decline due to voluntary unemployment because the non-monetary income (the value of leisure) is higher than the income (rent) of work. However, for the unemployed, the value of leisure is not so high as the minimum wage, because they are willing to work, but just no firms are willing to employ them. The non-monetary income earned by the unemployed in enjoying leisure must be lower than the income of work, so the rent of the whole society has indeed dropped due to the unemployment.

If the unemployed choose to peddle on the street, they are actually self-employed. There is no unemployment in this case, but if there were no minimum wage law, they would prefer to work in a firm. Although the equilibrium wage is lower than the minimum wage, it must be still higher than the income of peddling, otherwise they would have chosen to peddle instead of being employed by the firm. Just because there is no firm willing to employ them due to the minimum wage law, they have to be self-employed, and the income of peddling must be not only lower than the minimum wage, but also lower than the equilibrium wage determined by the market, which implies that the rent of the labors has decreased. In this case, the factor of labor has not encroached on the rent of other factors, but rent of it decreases, so the rent of the whole society still decreases due to the minimum wage law.

In a word, no matter whether the rent of the whole society decreases due to the rent-encroaching on other factors or on the taxpayers, or whether the rent of the labor decreases, the policies that raise the price of labor (wage) over the equilibrium (such as the minimum wage law) will cause the transaction cost to rise because of infringement of PPRs. And the decrease in the rent of the society is precisely rent dissipation, so rent-encroaching must lead to rent dissipation.

Section 4: The Causes of Unemployment

In this section, the causes of unemployment will be discussed in detail.

As mentioned in Lecture 9, there is only one direct cause of unemployment: the wage is higher than the equilibrium determined by the market, which is acknowledged by all economists. But why is the wage too high? Why is the wage not adjusted downwards? In other words, why is there "Wage Rigidity"?

Some policies of government are of course one of the most important reasons. (1) The minimum wage law directly abolishes the price adjustment mechanism in the labor market to eliminate unemployment by lowering wages. There are also policies such as requiring firms to bear part of social insurance benefits, which objectively force them to raise wages, thus cause them to reduce employment or even dismiss employees. (2) The trade unions have administrative monopoly power, and they strive to raise the wages at the expense of decreasing the employment. (3) The social welfare objectively encourages people not to work. If welfare benefits are higher than wages, self-interested people naturally choose not to work, which is logically similar to the choice of voluntary unemployment, but different in that the voluntary employed enjoy leisure on their own money, while the welfare recipients on the taxpayers' money. Moreover, government will not regard voluntary unemployment as unemployment, while the welfare recipients need to prove that they are unemployed.

Besides the government policies that will cause unemployment, there are other reasons for wage rigidity. The "Efficiency Wage Theory" argues that it is the firm that deliberately raises the wage above the equilibrium, because on the one hand high wage will make the employees devote to working in a better state, and on the other hand, it will create unemployment that frightens the employees not to shirk. This explanation seems to be reasonable and is logically similar to the strategy that the iPhone is deliberately priced lower than the equilibrium, which creates the queuing of buyers with advertising effect.

However, the efficiency wage theory is refuted by facts. When there is economic recession, there will be an increase in unemployment, because the difference between the wage and the equilibrium has increased. Economic recession will cause the demand for labor decrease, so the equilibrium of wage will also decrease as long as the supply of labor is unchanged. Even if the wage remains unchanged, the difference between the wage and the equilibrium will increase. According to the efficiency wage theory, it is enough that the wage is higher than the equilibrium, and there is no need to increase the difference between them in economic recession. In other words, the efficiency wage theory can explain the unemployment caused by the wage higher than the equilibrium, but it cannot explain the increase in unemployment caused by wage rigidity (the wage cannot fall) in economic recession.

The correct explanation still relies on the constraint that there is transaction cost (mainly information cost). Unemployment is usually divided into the following kinds: frictional unemployment, structural unemployment and cyclical unemployment.

Frictional unemployment occurs when one quits the old job and looks for a new one because he needs to take some time to find a new job. This kind of unemployment is obviously caused by information cost. Frictional unemployment will not last for a long time. As time goes by, the information cost will drop and this kind of unemployment will disappear. In fact, frictional unemployment is not an economic problem that government is worried about. In the market there are job-hunting agencies and headhunters to reduce the information cost.

Structural unemployment refers to the phenomenon that there are vacant positions and the unemployed at the same time. For example, there are many vacant positions in less developed areas, but few are willing to apply for those jobs, and most unemployed prefer to stay in developed areas. Of course, as long as the wage in less developed areas is high enough, this kind of regional structural unemployment can be eliminated. However, another kind of structural unemployment is more difficult to eliminate. There are many vacant positions in sunrise industries like IT, but the unemployed often appear in declining and shrinking sunset industries. Those unemployed do not have the appropriate skills to engage in new occupations. Although the wage of the vacant positions is very high, they are just beyond their reach. This kind of unemployment is caused by the adjustment of industrial structure. Re-employment training can solve the problems to some extent, but for the old unemployed who are difficult to learn new skills, they can only accept the wage reduction to stay in the old industry or turn to the similar industries with low wage but low skill requirement. In any case, as long as the wage can be flexibly lowered, those with low skill will not lose their jobs. Structural unemployment is due to the difficulty of matching the location or skill of the labor and the job, so the reason for it still lies in transaction cost.

The so-called cyclical unemployment is the concern of the government's macroeconomic policies. As has analyzed in Lecture 19, as long as there is no problem with monetary policy, there will be no business cycle. And even if there is business cycle, as long as there is no interference of the government policies on the labor market mentioned above so that the wage can be flexibly lowered, there will not be serious unemployment for long.

As mentioned in Lecture 19, gold had been used as money (gold standard) for thousands of years and had done a good job. However, strictly speaking, the gold standard is not the best monetary system because gold is not a good. It had done a good job as money just because it cannot be easily overprinted. In the long ancient time, the economic growth was very slow, and the gold supply also increased very slowly, so the growth rate of them could match with each other. However, since the industrial revolution, the economic growth rate has been significantly faster than that of gold, which implies that the gold standard will cause deflation. In fact, there was indeed severe deflation in Britain during the industrial revolution. But why was there not as serious unemployment and economic recession in Britain during the industrial revolution as that in the US during the Great Depression? Instead, during the industrial revolution, there was rapid economic growth in Britain. The reason is simple that there were no "evil laws" such as the minimum wage law, trade unions and social welfare that interfered with the labor market. Deflation will cause a drop in the prices in the product market, but the prices in factor market (including wage of labor) also drop, so firms can keep their profits and continue to operate. What is more, with the technological progress brought about by the industrial revolution, firms can obtain more output with less input, which is also helpful to fight against deflation.

There was a similar case in China from 1996 to 2000. There was no minimum wage law and very little social welfare in China, and the contract of piece-rate and share contract (dividends, bonuses) were widely used, which could automatically lower wages along with the economic recession with the greatest flexibility. As a result, deflation did not cause long-term unemployment, nor did it cause serious economic recession. Instead, it was a helpful pressure for firms to fight against deflation by improving quality and technology of products. In fact, the quality of products had improved significantly during that period, which laid the foundation for the products "Made in China" to conquer the whole world, especially after entering the WTO in 2001.

As mentioned in Lecture 13, if all production factors are purchased or rented through sharing contracts, there will be no unemployment, because the income (wage) of labor will decrease along with the decrease of the firm, and there will be not any rigidity. In fact, the unemployment rate in Japan has been much lower than that of western developed countries, and the "lifetime

employment system" is attributed to. However, the correct explanation is that in Japan's wage structure, there is a very low fixed wage as the base salary, and the rest is mostly performance-related bonus that is essentially sharing wage. When the income of the firm declines, the wage of the employees can be flexibly lowered, so there is little need to dismiss the employees. In other words, the lifetime employment system is not the cause but the result of low unemployment. This explanation can be tested by the fact that since 1990 when Japan's economy has been in a long-term recession, the income of some firms dropped sharply, even below the base salary (fixed wage), so there were a large number of dismissals, resulting in an increase in the unemployment rate. The lifetime employment system was easily abandoned. If all wages were sharing wages, the firms would not fire the employees even if the income dropped to zero, because they then used the labor for free. Of course, it was more likely that the employees would choose job-hopping.

The contract of piece-wage will neither cause unemployment. Under the contract of piece-wage, the factor market is inseparable from the product market, so the information of the price paid by final consumers is directly transmitted to the self-employed through the middlemen. As the price of product falls, the piece-wage must follow. If the self-employed do not accept the wage cut, the products cannot be sold out, not because the middlemen refuse to buy, but the consumers refuse to buy. What is more, the self-employed work for themselves, there cannot be unemployment.

It is the contract of time wage that will cause wage rigidity and unemployment. What the manager really wants is not the working hours of the employees, but the outputs produced during the working hours. Just because the transaction cost of directly measurement is too high, time is entrusted to indirectly measure the products. If the productivity of the firm declines, which indirectly implies that the time value of labor also declines, the time wage should be reduced. However, it is difficult to persuade the employees to accept the wage reduction precisely because it is indirect implication. Employees will argue that they work as hard and efficient as before, which means the quality of labor has not changed, why should the time wage be reduced just due to the economic recession? It sounds very reasonable, but the firms can argue similarly. The quality of the products has not changed, but now there is economic recession, and the demand for the products decreases, so the price has to be reduced.

The price of production factor is determined by that of product, which is always true even if the factor market is separated from the product market. When there is economic recession, the pressure of price reduction in the product market will be transmitted to the factor market. Just when the factor market is not separable from the product market, the suppliers of production factor and that of product are either the same individuals, or under the pressure directly, so the root of the price (or wage) reduction is clear and are easily accepted. By the contrast, when the factor market is separated from the product market, which means the contract of wage is time wage, the suppliers of product are directly under the pressure of price reduction, while the suppliers of production factor (employees) are indirectly under the pressure, so the root of the wage reduction is not clear, and the employees will tend to refuse to accept it. Obviously, it implies that the information cost in the latter case is higher than that in the former case.

Furthermore, time wage is fixed wage that cannot be adjusted in principle during the contract term, so it cannot be lowered along with the decline in the income of the firm during the contract term, thus forming wage rigidity. Of course, the contract can be terminated in advance, so in reality if the transaction cost of terminating the contract is lower than that of negotiating wage adjustment, the firm will prefer dismissal to wage reduction. The question then becomes: why do employees not accept wage cut but rather dismissal? It is logically similar to this question: why do the unemployed not accept a lower-paid job but rather remain unemployed? Social benefits such as unemployment benefits provided by government will certainly encourage people to choose unemployment instead of accepting wage cut (especially if the wage is reduced to be lower than unemployment benefits), but the cause of government intervention is not taken into account for the moment.

The reason is still the information cost. In the economic recession, even if the unemployed know that their wages need to be reduced, the problem is that they do not know how much to reduce, which means the information about equilibrium wage is sometimes not so clear in reality. If those with similar skills get the same wage, this information is not difficult to know. The unemployed will know clearly how much they should get when they observe the wage of those who with similar skills are still employed. It is logically similar that a seller of vegetables observes how much the other sellers price their vegetables in the market. However, if for some reasons those with similar skills get very different wages, it will be very difficult for the unemployed to judge what the equilibrium wage is.

Why do people with similar skills get very different wages? Firstly, those like civil servants seldom lose their jobs, and neither is there wage cut due to the economic recession, which makes their wages out of line with their skills. What is more, when there are those like civil servants who do not reduce wages during the economic recession, the others have to reduce more wages to achieve the overall equilibrium. For example, there needs to be a 10% of wage cut due to the economic recession, but 1% of the labor force are civil servants and will not reduce their wages, so the remaining 99% have to reduce their wages by $10.1\% (= \frac{10\%}{99\%})$ instead of 10%, which means the 99% have to share the part of wage cut of the 1%.

Secondly, as mentioned in Lecture 9 of the analysis on the minimum wage law, when there is economic recession, the firms will not immediately fire the employees, but cut the recruitment of new staffs and reduce wages in disguised form by increasing labor intensity for the employees. The wages of the employees seem to have not dropped, which will mislead the unemployed to think that they can find the same jobs as before, but in fact they are difficult to do so because the firms have cut the recruitment of new staffs. If government adopts the policies such as "dismissal compensation" that increase the transaction cost of dismissal, the firms will become more prefer to cutting the recruitment of new staffs rather than dismissing the employees, so the unemployed will be more seriously misled, implying an increase in information cost.

In a word, due to the wrong policies of government and some other constraints, those with similar skills have different wage rigidity, which increases the variance of wages and causes an increase in the information cost for the unemployed to correctly judge the equilibrium wage, and it in turn increases wage rigidity and causes more unemployment.

The definition of unemployment has been discussed in Lecture 9, which concludes that the only direct reason to unemployment is that the price (wage) of labor is too high. But why is the price of labor too high and not reduced? Besides the "evil policies" such as the minimum wage law and trade unions that directly prohibit the decline of wages, the other important reason is the information cost on the equilibrium of wages.

In fact, it is logically similar for the "unemployment" of other production factors. For example, it is well known that it takes a rather long time for a house to be sold out. The house is in idle before it is successfully sold out, which can also be regarded as "unemployment". Because the quality of a house is relatively complicated, including the quality of the building, design and even the location. Therefore, the information on whether the qualities of different houses are the same and whether the prices can be comparable to determine the equilibrium prices is rather higher than that of other goods. In addition, people are willing to spend more time searching for relevant information because the price of a house is usually high enough to burden very high information cost.

Turing from a house to a stock, there is no stock that cannot be sold out, but only the sharp drop in the stock price, which implies that as long as there is no difficulty in lowering the price, any good can be sold out.

Section 5: Natural Rate of Unemployment

As mentioned in Lecture 19, even Friedman, the head of the monetary school, advocates a moderate inflation is the lubricant of the economy. In fact, there is no function of lubricant of the economy for inflation in essence, but why do western governments generally like to have a moderate inflation? If there is minimum wage law, trade unions, social welfare and other factors that cause wage rigidity, a moderate inflation can help to secretly reduce the real wages without adjusting the nominal wages, thus effectively reducing unemployment and stimulating the economy.

The price in the theory of supply and demand is real price (relative price), but what can be observed in reality is monetary price (nominal price), so the equilibrium wage refers to the equilibrium of real wage instead of nominal wage. The real wage is the nominal wage divided by the price level ($w = \frac{W}{P}$). If the nominal wage (W) cannot be cut under the constraint such as minimum wage law, the real wage (w) can be cut to be closer to the equilibrium by increasing the price level (P), thus reducing the unemployment rate.

Obviously, the government makes use of the money illusion. However, illusion does not last long. Sooner or later, labor will find that there is inflation, and they will require the nominal wages to be raised based on the inflation rate. As a result, the decrease in unemployment caused by the drop in the real wages closer to the equilibrium will rebound again. Although the effect does not last long, the western governments are never bored with playing this trick, because the political system of democratic election encourages short-term behaviors. As long as the economy improves during the election and helps the politicians to be elected, it does not matter that the economy will return to the original state a little while later. With CPI, a vague and time-lagged measurement of inflation, it is easier for government to cause money illusion for a while even though many western workers have become smart and required in their wage contracts that the nominal wages must be adjusted based on the inflation rate.

In macroeconomics of MSE, there is a "Phillips Curve" that is used to show the substitution relationship between the unemployment rate and the inflation rate (the unemployment rate will be lower when inflation rate is higher, and vice versa). With the horizontal axis representing the unemployment rate and the vertical axis representing the inflation rate, the Phillips curve is sloping downwards as shown in Fig.20-4. It was originally only an empirical curve, proposed by Phillips, the economist in New Zealand, in 1958 in his article "The Relationship between the Unemployment and the Rate of Change in Monetary Wages in Britain from 1861 to 1957", based on the relevant statistical data. Later, Paul Samuelson (1915-2009) and Solow (R. M. Solow, 1924-) replaced the rate of change in monetary wages in the original Phillips curve with the growth rate of price. In macroeconomics, this curve is used to illustrate that inflation is driven by the rise in the cost of wage, which is totally nonsense! According to the monetary theory, there can be no inflation if money is not overprinted, so the cause of inflation cannot be the increase in wages. Without increasing money supply, even if the evil policies such as minimum wage law lead to a rise in wages, firms will only choose to lay off the employees to save cost instead of raising the price to cover the cost. The correct explanation for Phillips curve is the above analysis: when there is inflation, the real wage will fall to be closer to equilibrium wage, thus decreasing unemployment.

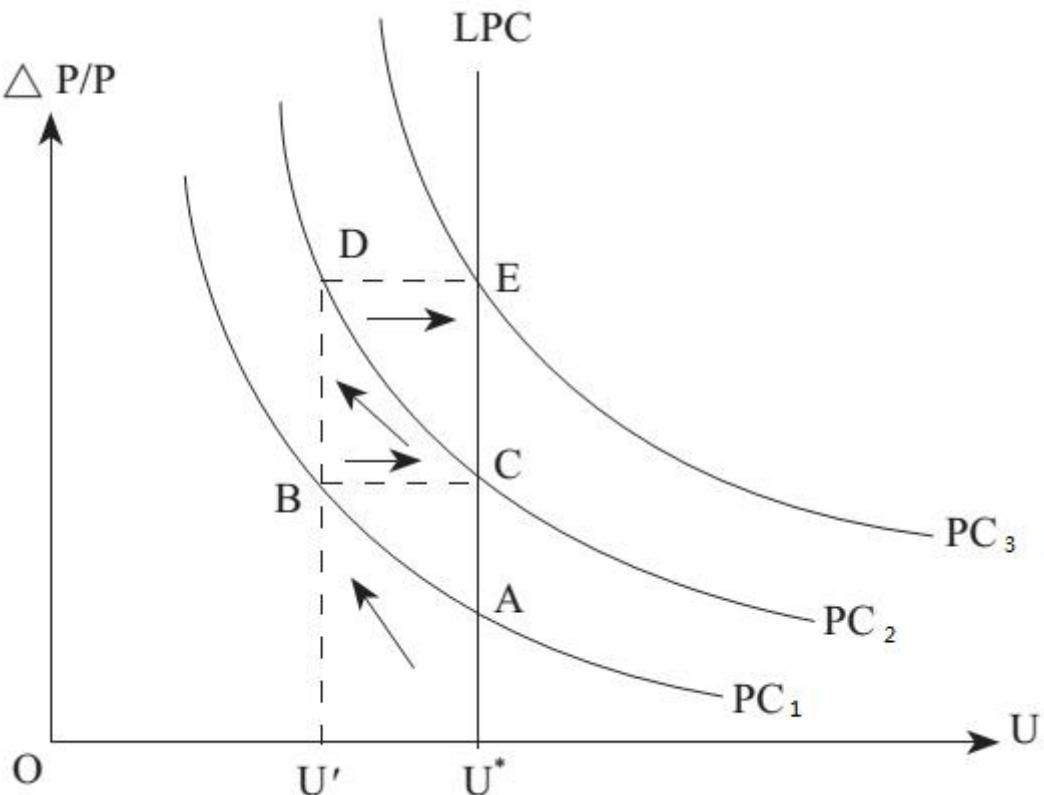


Figure 20-4

Later, Friedman put forward the “Expectations-augmented Phillips Curve” where the labor will know there is inflation in long run and requires adjusting the nominal wages based on the inflation rate. Once the nominal wage is adjusted based on the inflation rate, the real wage will return to the same as that before inflation, so does the unemployment rate, which has been shown in Fig.20-4 (from A to B then to C). Thus, the unemployment rate does not fall, but the inflation rate has risen. Therefore, the long-term Phillips curve (LPC) is a vertical line, different from the short-term curve sloping downwards, which means there is no substitution relationship between inflation and unemployment.

The intersection of the long-term Phillips curve and the horizontal axis is named by Friedman as the “natural rate of unemployment” (U^* in Fig.20-4). The definition of natural rate of unemployment in MSE is the equilibrium unemployment rate without the influence of monetary factors on supply and demand in the labor market and the product market. In other words, it is the unemployment rate under the condition of full employment, and it is the sum of frictional and structural unemployment rates”. However, this definition is also totally nonsense. If the labor market and the product market can be free from interference, how can be there unemployment? If it refers to the sum of frictional and structural unemployment rates, how can it be reduced by a moderate inflation? How can more money supply make the unemployed without IT skills find jobs in the IT industry?

The correct definition of the rate of natural unemployment should be the unemployment rate that cannot be eliminated fundamentally by inflation. The causes of it cannot be frictional or structural, but the institutional factors that prevent the market mechanism working and lead to wage rigidity, such as minimum wage law, trade unions and social welfare. Although the information cost will also cause unemployment, without the government interference, it will gradually drop over time, and the unemployment rate will also drop. Only the institutional factors can cause long-term unemployment.

According to the monetary theory, the monetary policy cannot adjust the economy but cause economic fluctuation or business cycle. Nor can it eliminate unemployment but cheat the labor for a short time with the money illusion. However, the labor is not foolish, so the final result is just inflation.

The unemployment caused by the institutional rigidity stemmed from evil policies can only be eliminated by institutional reform. For example, before the reform and opening up policy began in 1978, there were a large number of "unemployed youth" and also even more hidden unemployed in the state-owned enterprises because they actually did not need so many employees. There was such a joke: three workers were needed to fill a hole on the wall, one was carrying a bucket of cement paste, one was carrying a tool for filling the hole, and another was pointing at the hole with his finger! China transformed from the planned economy to the market economy, breaking the iron bowls of employees in state-owned enterprises. Although the employees in state-owned enterprises had experienced the pains of layoff, the rapid economic growth created a large number of jobs for them, which quickly eliminated the unemployment and did not have the adverse consequence of inflation.

As for the unemployment caused by information cost, because the information cost will gradually decrease over time, and there will also be information experts in the market to help to reduce it in order to earn the rent of information. If government wants to help, it should try to effectively reduce the information cost. However, in reality it is much more usual that what government does is to increase the information cost.

Lecture 21: The Critique on Keynes's Theory of National Income Determination

In Lecture 6, we have introduced how Keynes gained his fame by explaining the causes of Great Depression. He put forward the theory of national income determination, the basic conclusion of which is that the free market will cause Aggregate Demand to be less than Aggregate Supply in the economy, so government needs to adopt economic policies to increase aggregate demand to make it equal to aggregate supply and reach the equilibrium of national income.

However, we have already analyzed in detail that the real cause of overproduction is the wrong monetary policy. We have also explained that economic fluctuation (business cycle) is not the result of free market, but the wrong monetary policy again and some government policies that interfere with or even abolish the market mechanism that adjusting the price to equalize supply and demand. Precisely because the Keynes's theory has still influenced government policies, it is necessary to introduce it and point out the mistakes one by one.

Section 1: The Statistical Index of National Income: GDP

Before introducing and criticizing Keynes's theory, we will first introduce the statistical index widely used to measure national income -- Gross Domestic Product (GDP). This statistical index is based on Keynes's theory of national income determination, and although the theory is wrong, GDP has been widely used to measure national income in practice.

GDP is defined as the market value of all final goods and services produced within a country or region in a given period of time (usually a quarter or a year). A detailed explanation of this definition is as follows.

(1) GDP is used to measure the national income in a given period of time, which means it is a flow. In Lecture 18, the distinction between flow and stock in statistics has been introduced. National income is the sum of the income of all within a country or region. Since income is a flow, national income is of course also a flow. The period for measuring this flow is usually one quarter or one year, so in financial reports there are usually the data of quarterly GDP or annual GDP.

(2) GDP is the market value, which means the monetary value of market transactions. It is the market value instead of quantity because different goods have different units of quantity. If the quantity is not multiplied by the monetary price, it cannot be summed up.

However, there will be some problems. Firstly, the monetary price is nominal with inflation, so the increase in GDP is not only due to the increase in the quantity of goods, but also includes the factor of price rise, and the price rise may be caused by various factors such as the improvement of quality, changes in supply and demand, inflation, etc.

Therefore, GDP is also divided into nominal GDP and real GDP. The former is calculated by multiplying the current output by the current price of goods, while the latter is calculated by multiplying the current output by the base price of goods. For example, if 1978 is taken as the base period, the nominal GDP in 2011 is calculated by multiplying the price in 2011 by the output also in 2011, while the real GDP is calculated by multiplying the price in 1978 by the output in 2011, which can eliminate the impact of the price rise from 1978 to 2011. The GDP of China is nominal one, but the growth rate of GDP is real one.

Just as there are many problems with CPI (a statistical index for measuring inflation), so are real GDP. The general rise in prices caused by the general improvement in quality will be wrongly regarded as inflation, and these two indexes cannot handle the emergence of new products.

Secondly, if a good or service is not transacted through the market, GDP cannot measure the value of it. The most typical example is household service. If a wife chooses not to work and stays at home to take care of her husband and children, the GDP of the service she provides for the family cannot be measured. However, if she chooses to work and hires a nanny to do housework and take care of her children, the nanny provides the same service as the housewife in the previous case, but the service is transacted through the market, so GDP can measure the value of it. Thus, there is the same household service in these two cases, but the GDP in the latter will be higher than that in the former.

Besides, illegal and black market transactions can neither be measured by GDP. There are huge gray economies in many developing countries, so the data of GDP must be underestimated.

(3) GDP measures the value of final goods. There are intermediate products in the process of production, but GDP does not measure the value of them, otherwise there will be the problem of double calculation.

For example, the price of a piece of bread is RMB1, and there are intermediate products of wheat and flour. Suppose the value the wheat needed in the production of this piece of bread is RMB0.5, and that of flour is RMB0.7. GDP only include the value of the final bread (RMB1), instead of the sum of the value of bread, wheat, and flour. Because the value of flour (RMB0.7) has included that of wheat (RMB0.5), while the value of bread (RMB1) has included that of flour, if GDP include the value of the intermediate products, there will be three times of the value of wheat and twice of the value of flour. Of course, some flour is bought by the final consumers as flavoring, they are final products instead of intermediate products, and should be included in GDP.

(4) GDP measures the value of finished products instead of products sold out. GDP is a flow for a quarter or a year, so some products have not been sold out, which means they are inventory, and should be measured in GDP. If some goods were produced in the previous quarter or year, but sold out in this quarter or year, they cannot be measured in the GDP of this quarter or year.

What is more, the value of newly-built houses should be measured in GDP, but that of second-hand houses should not, while the transaction costs such as commissions paid for the transactions of second-hand house should be measured because it is the services provided in this quarter or this year.

(5) Both products and services should be measured in GDP. Products are tangible goods, while services are intangible goods, and both are valuable. The statistical index "aggregate social product" used to measure national income in China before is based on the "system of material product balances" established by Marx's economics. As the name implies, this system only measures "material products", not including intangible services. There had been a joke: For a train, the transportation cost of goods is measured as national income, but that of human is not.

For the statistical practice in reality, there are three approaches for measuring GDP: expenditure approach, income approach and production approach. In theory, these three approaches should get the same data, but there are inevitably statistical errors. The expenditure approach is based on Keynes's theory of national income determination, so it is the basic approach among the three, and the data of GDP released by government is also based on it. Here, only this approach is introduced.

The formula of GDP by the expenditure approach is: $GDP = C + I + G + (X - M)$

"C" refers to residents' consumption expenditure, including expenditure on various durable and non-durable consumer goods and services, but excluding that on housing construction that is included in I (investment).

"I" refers to the investment expenditure of enterprises. The term of "investment" here is not the general concept mentioned in Lecture 18, but refers specifically to the behavior of an enterprise investing money into doing business. Moreover, it only includes the fixed-asset investment and inventory investment. As mentioned earlier, if the products have not been sold out during the accounting period, they are in the form of inventory and should be measured in the current GDP. From another perspective, investment can also be divided into net investment and replacement investment that is used to make up for depreciation.

"G" refers to government purchase including all levels of government spending on products and services. Attention: it does not include government's Transfer Payment that is the government's expenditure on social welfare, social insurance, unemployment relief, poverty subsidies, aged security, health care, subsidies for agriculture and other industries, etc. Transfer payment is only the unilateral monetary expenditure and does not exchange for products or services. The

"expenditure approach" measures national output or income from the perspective of expenditure (the price of a good is income for sellers, but expenditure for buyers). If the expenditure does not exchange for corresponding output, it should not be measured. In other words, transfer payment does not increase a country's income or wealth, but only redistributes income or wealth.

"X - M" is the net export because X refers to export and M refers to import. Export is the good produced domestically but consumed by foreigners, so it should be measured in GDP. By contrast, import is the good produced by foreigners but consumed domestically. This part of consumption has been measured in C, but it is not actually domestic product, so it should be deducted here.

In a word, C is the expenditure of individuals or households, I is that of enterprises, G is that of government, and X - M is the net expenditure of foreigners on domestic products. The expenditures of the above four sectors have purchased all the products in an economy. Financial reports often talk about the three carriages driving economic growth, which refer to the expenditures of consumption, investment and net exports, because these expenditures are determined by the market and can objectively reflect a country's productivity and competitiveness.

By the way, from the above introduction of GDP, it can be concluded that national income is the same as national output, because the output will become income after sale.

Section 2: Is Saving Different From Investment?

From now on, we will introduce and criticize Keynes's theory of national income determination in detail.

The formula of GDP is based on Keynes's theory of national income determination that is a model of four-sector economy. For the sake of simplicity, we will first introduce the simple model of two-sector economy that there are only the sectors of households and enterprises without the sectors of government and foreign trade.

What is more, GDP measures the actual national income, which means it is about the historical data after production and consumption. By contrast, the theory of national income determination analyzes the equilibrium national income that refers to the national income when the planned aggregate demand and aggregate supply are equal before production and consumption. In other words, the former is a fact and the latter is an intention, similar to the difference between the trading volume (quantity purchased and sales volume) and the equilibrium quantity (quantity demanded and quantity supplied) in the theory of supply and demand.

The Aggregate Expenditure or aggregate demand of the two-sector economy is $E = c + i$ where E refers to aggregate expenditure. Keynes proposes the consumption function $c = c(y)$ ¹ that is supposed to be a linear function $c = a + by$ for simplicity. Here "a" refers to autonomous consumption that has no relationship with national income (y), because even if one does not have any income, he still has some consumption expenditures such as food to survive. And "by" refers to induced consumption that increases along with the increase in " y ". Keynes calls b "Marginal Propensity to Consume" (MPC), and the definition formula is $MPC = \frac{dc}{dy}$, so it is the coefficient of "b" in the above linear consumption function. The value range of MPC is between 0 and 1, which implies that as income increases, one will increase consumption, but the growth rate of consumption is lower than that of income, because he will not spend all the increased income to increase consumption, but will only spend some of it and save the rest.

As for investment (i), Keynes believes that it is decided by enterprises and has no relationship with " y ", so it is a constant.

The aggregate income or aggregate output of the two-sector economy is $y = c + s$, because the income obtained from the goods after sale is either spent or saved. When all planned outputs are purchased by planned expenditures, the product market will be in equilibrium: $E = y$ or $c + i = c + s$, arranged as $i = s$. Thus, national income is in equilibrium when the planned investment is equal to the planned saving.

Attention: in reality, all investments come from savings, so the actual investment must be equal to the actual saving, which is logically similar to the fact that the quantity purchased must be equal to the sales volume. However, Keynes argues that here " i " refers to planned investment and " s " refers to planned saving, and both are intentions, so they are not necessarily equal, and are only so when national income is in equilibrium.

However, there are problems in Keynes's argument. (1) Intention is not a fact, so how is this argument tested? In order to test it, there must be some objective facts induced from the intention that is tested indirectly by observing these facts. Keynes has not said anything about how to test the intended investment and saving. (2) As mentioned in Lecture 7, unless against the postulation of self-interest, the world is always in equilibrium. If there were disequilibrium, in the analysis there must be some constraints that have been ignored. (3) As mentioned in Lecture 18,

there is no difference between saving and investment. The remaining part of the present income after deducted the consumption is "saving", while the remaining part of the present income that is not consumed now but left for future consumption is "investment". They are the same, and just the former is from the perspective of present, while the latter is from the perspective of future.

What is more, Keynes regards consumption as the behavior of increasing national income (aggregate expenditure or aggregate demand), but saving as that of reducing national income, thus putting forward the so-called "Paradox of thrift". The behavior of reducing consumption or increasing saving will reduce national income and cause economic recession, so it is evil. By the contrast, the behavior of increasing consumption or reducing saving will increase national income and bring about economic boom, so it is good.

Keynes also points out that MPC is less than 1, which means one's income obtained from production will not be all spent but some of it will be saved. He also put forward the psychological "Law of Diminishing Marginal Propensity to Consume", which means as income increases, one's consumption will increase less and less, or he will save more and more. Without government interference, aggregate expenditure (aggregate demand) will be less than aggregate output (aggregate supply), which implies overproduction, resulting in economic recession.

However, saving is not different from investment. Moreover, now one usually deposits money in banks, and banks will lend the money to others in consumption (such as buying a car with mortgage loans) or in investment (such as borrowing money for an enterprise's operation), so how can aggregate demand or expenditure (consumption and investment) of a society decrease due to the rise in saving?

What if one keeps the money at home instead of depositing it in banks? Keynes puts forward the "liquidity preference" that argues that people are willing to hold cash instead of other non-cash assets because cash has liquidity that are preferred. He further argues that liquidity preference implies insufficient consumption and investment.

As mentioned in Lecture 19, banks usually lend out the money saved by depositors, but if they are extremely pessimistic about the economic prospect, they will be reluctant to lend. Even if one does not keep the money at home, banks also keep it in the warehouse and do not lend it out, so the money will not turn into consumption or investment of others. The investment seems to be less than and different from saving in this case, and excess savings seem to reduce aggregate demand. However, cash is only money, not goods. The goods or resources it can exchange for are not kept from production. From the perspective real goods, the investment of the whole society has not decreased.

According to the monetary theory, the above behavior of hoarding money will lead to the leakage of monetary base from the banking system and the decrease of deposit money created by commercial banks. As a result, the effective money supply in the economy will decrease, which will cause deflation. In other words, if people keep the money at home or banks are reluctant to lend out the deposits, there may be economic recession, but the reason is not that the decrease in consumption and the increase in saving will cause reduction in national income, but that money illusion caused by the unstable value of money misleads people to make wrong decisions. Under deflation, producers observe the price fall and think their products are not popular in the market, so they choose to reduce production, resulting in economic decline.

There is a story in the South Korean TV drama "Uprising" that should be based on the real historic event. The rich in ancient Korea liked to bury the copper coins in their backyards, resulting in serious "money shortage" (the ancient term of deflation). The hero of this TV drama was the richest in ancient Korea, and borrowed money from the rich, so that they dig out the copper coins and give to him, and then he lent the money to those who needed it. Since he was the richest, the rich all believed that he could repay the loans, and were willing to lend the money to him at a

lower interest rate. He loaned the money out at a higher interest rate and earned a profit from the difference. With more money flowing into economy, it was easier for those who borrowed money from the hero to succeed in doing business, so they could repay the money that was returned to the rich through the hero, and the rich became more willing to lend the money to the hero... In such a virtuous circle, the hero successfully solved the problem of "money shortage". Obviously, the hero in ancient Korea of this TV drama had played the role of commercial banks in modern society.

There is another solution that the central bank can issue more money to offset the decrease in effective money supply caused by the behavior of hoarding money. The monetary system pegging money to the price index of a basket of tradable goods introduced in Lecture 19 can monitor in time and accurately adjust the money supply to eliminate any inflation or deflation.

And what if one hoards real goods instead of money? As the saying goes, "save for a rainy day", one hoards grains against the famine, and the grains are neither consumed nor sown in the fields. Will this behavior of hoarding grains or saving result in aggregate demand (consumption and investment) less than aggregate supply (aggregate output)? The answer is NO! This so-called saving is actually consumption, just what not the consumption of the use value of grains to fill one's stomach, but that of the insurance that there are enough grains to fill one's stomach in case of famine. In other words, what is bought by hoarding grains is insurance.

Obviously, it is the consumption expenditure caused by transaction cost. Only when there is risk (measured by information cost), people need to buy insurance against risks. If there were no risk of famine, all grains would be used as food in consumption or as seeds in investment (production) and the future output would be higher. When the risk rises sharply due to some reasons, people will spend a large amount of current income to buy insurance that is also a kind of consumption expenditure, thus greatly reducing investment expenditure and causing future output to fall which implies an economic recession (negative output growth). There is no difference from the case that one changes the consumption pattern for some reason, from thrift to extravagance (perhaps because he knows that he is terminally ill and will soon die), so he spends all income on consumption expenditure, without savings turned into investment expenditure, which causes his future income to fall. Obviously, in both cases, the saving decreases rather than increases, which proves the true cause of the reduction in national income is not that the increase in saving will lead to aggregate demand less than aggregate supply, but that the proportion of income allocated to saving is reduced, thus reducing the investment that is the same as saving, resulting in a decline in future output.

In fact, some consumption or investment expenditure may be not helpful or even harmful to the increase in future output, causing the growth rate of output to slow down or even become negative. As mentioned in Lecture 12, consumption can be regarded as a behavior to produce labor as a production factor. And as mentioned in Lecture 18, consumption is also investment, only from different perspectives. Not all consumption or investment expenditures will increase the future output. Among them, the most typical example is the consumption of drug abuse that harms one's physical and mental health. If there are many who have addicted to drugs in a country, such as the case of China in late Qing Dynasty, the economy is sure to decline. Obviously, this economic recession does not caused by the increase in saving resulting in aggregate demand less than aggregate supply.

Similarly, some investment expenditures are not helpful or even harmful to the future output. As mentioned in Lecture 19, different assets have different liquidity. The illiquid assets can also attract investors because they have better profitability as compensation. If investors are very pessimistic about the future, which means the risk increases, the profitability after deducting risk premium will decrease, so few are willing to invest in illiquid assets. Such illiquid assets are actually more helpful to increasing the future output than the liquid assets. For example, cash is

the most liquid, but if one only holds cash and does not use it to buy goods or other assets, there will be no profitability at all, so it will not lead to any increase in the future output. Obviously, with higher risk, one would rather abandon the illiquid assets with better profitability, which is a reduction in investment expenditure and actually an increase in consumption expenditure of buying insurance.

In other word, such investment that is not helpful to increasing the future output is in essence the consumption that is not helpful to the future output. Not just drug abuse is unhelpful to increasing the future output, as does insurance, but the cause is that there is transaction cost. Unless the transaction cost can be effectively reduced, people will still choose to buy insurance, which is precisely the correct explanation for liquidity preference. Keynes made mistake because he did not know the concept of transaction cost, which has again proved that ignoring the transaction cost is the causes of many mistakes in MSE.

There is another kind of investment that seems to be not helpful to the future output: the expenditure on jade, artwork, etc. that can be collectively referred to as collection. The non-monetary value of enjoyment of collection is consumption expenditure, while extra value above consumption is investment (saving). Collection is not helpful to increasing the future output, and will it lead to economic recession? The answer is NO! Collection is a store of wealth, which is logically similar to the function of money as a store of value. As has been analyzed earlier, the behavior of hoarding money will not reduce the quantity of goods in the real economy, so similarly collection is not money, and storing it will not even cause a decrease in money supply. There is not much difference to store wealth in collection and to deposit money in banks. What is the difference between those who sell out collection to spend or invest and banks that lend out the money to those who need it to spend or invest?

To sum up, saving is not different from investment, so aggregate demand cannot be less than aggregate supply. There are two causes for economic recession (negative growth of output). One is the wrong monetary policy misleading people. The other is the rising transaction cost (risk) that makes people spend more income on buying insurance that is actually a kind of consumption expenditure unhelpful to increasing the future output. In other words, besides the wrong monetary policy, the increase in transaction cost is the real cause for economic recession. Think more deeply, the negative effect of wrong monetary policy is precisely leading to the increase in transaction cost because it is the transaction cost of monetary system due to the instability of monetary value.

In the next lecture, we will explain in detail that it is precisely the change in the proportion of transaction cost in social wealth (national income) that is the key to determine the economic performance.

Keynes's analysis on the equilibrium condition of national income is completely wrong! Based on this wrong analysis, Keynes put forward the "paradox of thrift", arguing that too little consumption and too much saving will lead to insufficient aggregate demand and economic recession. Thus, he advocates more consumption and less saving, and criticizes the traditional virtue of thrift. However, if there is little saving, where can investment come from? Even Marx knows that expanded reproduction needs surplus products. Keynes's argument has also been refuted the facts that East Asian countries with high savings rates outperform economically other developing countries with low savings rates.

According to the correct analysis, if government wants prevent economic recession, it should adopt the monetary policy that stabilizes the monetary to prevent economic fluctuation, and should try to reduce transaction cost to reverse people's pessimistic expectation for the future. These are precisely the effective solutions.

1. The lowercase implies that all values have been adjusted for inflation, and they are real values instead of nominal ones. [↵](#)

Section 3: The Problem in the Theory of Multiplier

As has been analyzed in the previous section, it is wrong that Keynes's theory of national income determination thinks saving is different from investment, and the increase in saving will reduce aggregate demand, thus decreasing national income.

There are still more mistakes in Keynes's theory. Based on the wrong conclusions, Keynes believes that government should try to increase aggregate demand, stimulate consumption and encourage investment to fight against economic recession. He also puts forward the "Theory of Multiplier" to illustrate that the policy of increasing aggregate demand can increase several times more of national income. The key to this magic is that MPC is less than 1.

For example, suppose the investment (i) has increased by 100 units for some reason, according to the equilibrium formula of $y = c + i$, national income (y) has also increased by 100 units. However, do not forget the consumption function ($c = a + by$). When y increases by 100 units, consumption (c) also increases by " b " (100 b) units. Suppose $MPC = b = 0.8$, so c has increased by 80 units. And c is also a component of $y = c + i$, so y also increases by 80 units. Once y has increased by 80 units, according to $c = a + by$, it will cause c to increase by $0.8 \times 80 = 64$ units. The increase in consumption again leads to the increase in national income according to the equilibrium formula, and the increase in national income in turn leads to the increase in consumption according to the consumption function ... Such cycle continues, just like the monetary multiplier introduced in Lecture 19.

Let us arrange the increase in national income of each round. In the first round, the increase in investment (i) leads to the increase in national income (y) by 100 units. In the second round, the increase in consumption (c) leads to the increase in y by 80 units. In the third round, the increase in c leads to the increase in y by 64 units ... Obviously, it is an infinite decreasing geometric progression, and each value is 0.8 of the previous value, which is precisely MPC. Thus, how much has national income increased in the end? According to the summation formula of infinite decreasing geometric progression, the answer is $\frac{a_1}{1-q} = \frac{100}{1-0.8} = 500$. The initial investment (i) increases by 100 units, while the final national income (y) has increased by 500 units, which is 5 times the growth of investment and so is the multiplier. As the increase in national income stems from that in expenditure, it is called "expenditure multiplier", and the general formula is $\frac{1}{1-MPC}$. The above example is about the increase in investment expenditure, so it can also be called "investment multiplier". If it is the increase in consumption expenditure that leads to that in national income, it is called "consumption multiplier".

This theory of multiplier of Keynes seems like the magic of Midas touch. As long as a kind of expenditure increases a little, national income will increase by several times. Only with common sense, one can know it is impossible, but there have been so many believers for decades.

Furthermore, some even advocate an absurd theory of broken windows. The shop windows of a shop haven been broken by a thief, so the shopkeeper has to buy new glass to replace the broken windows, which means his expenditure has increased and the seller of glass has also increased his income. The seller of glass will spend part of the increased income, so the seller of the relevant good has also increased his income and will also spend part takes out part of the increased income...Such cycle continues, so national income of the whole economy will increase by several times in the end, which attributes to the thief who has broken the windows. If it were not for him, how would the shopkeeper increase expenditure on new glass, and how can there be

a series of cycle increases in income and consumption? It is so absurd that this theory of broken windows advocates that "Theft is justified."

There was a clear-minded person (not even an economist) long ago who has pointed out the mistake of the theory of broken windows. A reporter of New York Times has written a small book that easily overthrows this myth. The money the shopkeeper spends on new glass might have been intended to buy a pair of shoes, but with broken windows he is forced to divert the money to the new glass, so his planned expenditure on new shoes disappears. According to the theory of multiplier, the income of the seller of shoes has decreased, so he has to cut some of his expenditure, which has reduced the income the seller of the relevant good, and the latter also has to cut some of his expenditure... Such cycle continues, so national income of the whole economy will decrease by several times in the end, which also attributes to the thief who has broken the window. If it were not for him, how would the shopkeeper decrease expenditure on new shows, and how can there be a series of cycle decreases in income and consumption?

In other words, the advocates of the theory of broken windows have a pair of problematic eyes of only seeing part of the facts, because they only see the destruction of glass has led to an increase in some unplanned expenditure, but not see that some planned expenditure has also decreased as a result.

One may ask what if the shopkeeper has not planned to buy a new pair of shoes. If the money is not planned for expenditure, it will be saved. As mentioned in the previous section, saving will turn into investment. If it is deposited in a bank, the bank will lend it to those who need to consume or invest. In a word, from the perspective of the whole economy, the money will always be spent on someone's expenditure and will not disappear even if it is not spent on new glass.

Which one is larger, the increase in national income caused by the increase in unplanned consumption larger, or the decrease in national income caused by the decrease in planned consumption or investment? It seems to involve complicated calculations of the elasticity of demand, supply and output. In fact, there is an easy method to go round all the complexities. The thief has broken the windows, which is sure to lead to less wealth, so how can theft be justified to increase wealth? Of course, there is a Chinese idiom "Without destroying the old, one cannot build the new." But if there is a plan to destroy the old and then build the new, self-interested people will calculate the loss of destroying and gain of building to make sure the loss is less than gain. How can the unplanned loss caused by a thief happen to be less than the gain of being forced to buy new glass? If it were, the shopkeeper would have broken the windows himself and need not wait for the thief to "help".

According to the logic of the theory of broken windows, should government encourage instead of punish theft? Only a fool would believe such a ridiculous theory. However, there do be economists cheering and applauding the major disasters such as earthquake, believing that they can help to increase investment expenditure and thus stimulate economic growth. If disasters were really good for economy, government should just bomb out buildings during economic recession, then rebuild them, then bomb out them again, then rebuild them again ... then the economy would recover strongly. Bombing out buildings might be too troublesome, and government would better encourage people to commit suicide, then more babies could be born, and the cost of raising a child could be much more than rebuilding a house. When the children grow up, they should be encouraged to commit suicide again. According to this logic, terrorists and killers are contributors to economic growth!

Section 4: The Problem in Keynes's Consumption Function

There is another fatal weakness in Keynes's theory of multiplier that it depends heavily on the consumption function. The logic that the increase in expenditure can cause several times of increase in national income depends on the fact that consumption (c) not only determines national income according to the equilibrium formula, but is also determined by national income according to the consumption function. Only when the two affect each other can such a circular effect be formed. The question is: Is the consumption function in reality like that advocated by Keynes?

Friedman, a leading figure of the monetary school, has been introduced in Lecture 6, but his criticism of Keynes' theory is not limited to monetary theory. In 1957, he published the famous paper "A Theory of the Consumption Function" that contradicted Keynes's consumption function and put forward the "Permanent Income Hypothesis".

Friedman points out that Keynes's consumption function is a theory of short-term (current) consumption, which means income (y) and consumption (c) in Keynes's consumption function are current income (y_t) and current consumption (c_t). By contrast, Friedman's permanent income hypothesis holds that there are two parts in current income: permanent income (y_p) and temporary income (y'). Similarly, there are also two parts in current consumption: permanent consumption (c_p) and temporary consumption (c').

Permanent income refers to the stable income over a long period of time. It is calculated by multiplying the sum of the discounted value of expected future income by a stable interest rate, so it is stable and unchanged for each period. Permanent income is actually the "Annuity" introduced in Lecture 18. Similarly, permanent consumption refers to the stable consumption over a long period of time, and it is calculated by multiplying the sum of the discounted value of expected future consumption by a stable interest rate.

Because both permanent income and permanent consumption are stable, there is a definite functional relationship between them, namely $c_p = c(y_p)$. However, temporary income is occasional income that cannot be expected. The same is for temporary consumption. From the statistical perspective, temporary income and temporary consumption are essentially random variables, so there is no definite functional relationship between them. Obviously, they are profits introduced in Lecture 12 because they are unexpected. By contrast, permanent income and permanent consumption are both expected.

Attention: permanent income or consumption is stable, which is based on some expectations, but expectations can be adjusted along with the change in the constraints, so they are called "Adaptive Expectations". For example, one has won a lottery of 1 million, which cannot be expected in advance, so it is an unexpected profit that is temporary income. However, after it has happened, the winner knows for sure that his current wealth has increased by 1 million, which implies the constraint of wealth has changed. Suppose the stable interest rate is 10% per year, the winner can definitely expect that he has more income from the interest of 100,000 per year, so his permanent income will increase by 100,000. Therefore, that permanent income is stable does not mean it will never change, but only that it is the expectation based on the current constraints.

Thus, current income is $y_t = y_p + y'$, while current consumption is $c_t = c_p + c'$. Because temporary income (y') and temporary consumption (c') are random variables, current income (y_t) and current consumption (c_t) are also random variables. Keynes believes there is functional relationship between current income and current consumption, which is completely wrong!

Of course, Friedman's theory is only a "hypothesis". The core content of his famous paper "A Theory of the Consumption Function" is precisely about the test of this hypothesis of permanent income. Once published, this paper has been regarded as a classic work not only because it criticized Keynes, but it has fully demonstrated Friedman's profound ability as an economic master using statistical data to test theoretical hypotheses.

In fact, Friedman was a statistician before he became an economist. His work during the World War II was to design weapons and strategies and to analyze data from metallurgical experiments for the US army, so he was good at statistics. In the permanent-income hypothesis, there is the far-reaching influence of statistics.

For example, there are two parts in the current income (consumption): the fixed permanent income (consumption) and the uncertain temporary income (consumption), which is precisely the routine of statistics that in a regression equation there are two parts of the fixed functional variables and the uncertain random variables.

What is more, permanent income (consumption) is calculated by multiplying the sum of the discounted value of expected future income (consumption) by a fixed interest rate. The specific calculation is to give different weights to the expected income (consumption) at different times: the farther away from the present, the less accurate the expectation is, and the less influence on the current income (consumption), so the smaller the weight is. Obviously, from the statistical perspective, permanent income (consumption) is actually a weighted average.

Friedman's permanent-income hypothesis has a very important policy implication: it has negated the policy implication of Keynes's consumption function theory that the increase in (current) income can increase (current) consumption, thus stimulating economic growth. Only the economic policy can effectively increase permanent income, which means it can make people form the expectation that their incomes will increase in the future, will permanent consumption increase, thus increasing current consumption and effectively stimulating economic growth. However, the policies of stimulating aggregate demand advocated by Keynes can only increase temporary income and temporary consumption, and cannot increase current consumption at all, so the logic chain of multiplier effect is interrupted.

Section 5: The Pandora's Box of Fiscal Deficit

There are still more mistakes in the policy recommendations deduced from the wrong Keynes's theory of national income determination. In the previous sections, only the simplest model of two-sector economy, where there is no government, is analyzed. If extended to the model of three-sector economy, there will be government, so Keynes's theory has opened a terrible Pandora's Box by advocating increasing government expenditure to stimulate economic growth!

According to his wrong theory of multiplier, Keynes believes that government expenditure is similar to consumption and investment, which means an increase in government expenditure can bring about several times of increase in national income. Consumption and investment are both private decisions, and government just can encourage individuals and firms to increase expenditure, so the effect is uncontrollable. By the contrast, government has full decision right on government expenditure.

Keynes even advocated during the Great Depression that government should hire a group of labor to bury bottles in the ground, then dig them out, and bury them again ... Such meaningless work could be good to economy because it would make the employed labor have income that would be spent, so the sellers of the consumed goods would have income and increase consumption that would increase the income of another sellers ... Such cycle continues, so national income of the whole economy will increase by several times in the end. Obviously, it is logically similar to the proposal in section 3 of bombing out buildings, rebuilding them and then bombing out again ..., so it is similarly ridiculous.

Is government really so foolish to believe such an absurd policy recommendation? Of course not! If one believes government really believes in Keynes, it is he who is foolish! Do not forget that government spends the taxpayers' money instead of its own to increase government expenditure! In fact, government had long been eager to spend freely, but it had been constrained by the prudent fiscal discipline of "keep expenditures within the limits of income" advocated by classical economics which also believed that a small government should not interfere with the market. Well, now there is a guy named Keynes who announces that increasing government expenditure is good to stimulating economic growth! For such a so-called "scientific theory", it is no wonder that government thanks Keynes much and seems to believe in him as a god.

Wait a minute! The money spent by government comes from taxpayers. According to Keynes's theory of multiplier, although the increase in government expenditure can lead to several times of increase in national income, but the increase in tax will also lead to several times of DECREASE in national income, so the net effect on national income will depend on which multiplier is larger. And according to Keynes's theory of national income determination, in the case of three-sector economy, $y = c + i + g$, where consumption function is $c = a + by'$, and y' is "disposable income", the definition formula of which is $y' = y - t$ (t is tax). Thus, the government expenditure multiplier can be calculated to be $\frac{1}{1-b}$, where " b " is MPC, while the tax multiplier is $-\frac{b}{1-b}$ that is negative, which means the increase in tax will lead to the decrease in national income. In other words, if government increases tax to increase expenditure, the multiplier effect will be nullified because the sum of government expenditure multiplier and the tax multiplier is precisely $\frac{1}{1-b} - \frac{b}{1-b} = 1$

However, Keynes points out that the reason is that classical economics requires government to keep expenditures within the limits of income and allow no fiscal deficit (also known as "budget deficit"). If government is allowed to have fiscal deficit, which means the expenditure can be more than income, government can increase expenditure without increasing tax revenue, or even cut tax to reversely use the effect on national income of the tax multiplier to further simulate economic growth.

The question is: how can government spend more than the income? If one spends more than his income, he can borrow money to make up the gap. Similarly, if government has fiscal deficit, it can also borrow money, which is the issuance of treasury bonds. The money borrowed has to be repaid, and one should work hard to earn more future income to repay the debt, but what about government? The income of government comes from tax, so treasury bonds are repaid with future tax, unless government quietly prints more money to repay the debts, which will cause inflation or even hyperinflation. In other words, if government does not engage in inflation, the multiplier effect of tax on decreasing national income is just postponed if government increases expenditure by borrowing instead of increasing tax, which has not fundamentally eliminated the offset of the multiplier effect of tax on that of expenditure.

However, Keynes still thinks it does not matter. When there is economic recession, government should increase expenditure and cut tax to stimulate economic growth with budget deficit. And when there is economic prosperity, government can collect more taxes and spend less, so there will be fiscal surplus (budget surplus) that covers the deficit before. Classical economics is wrong to require government to keep budget balance every year. The correct approach should be to keep budget balance within an economic cycle.

This is precisely "fiscal policy". Together with the "monetary policy" mentioned in Lecture 19, it is one of the Keynes's two major policy tools in regulating the economy. Expansionary fiscal policy and expansionary monetary policy are used to stimulate economic growth during economic recession, while tight fiscal policy and tight-money policy are used to cool the overheated economy during economic prosperity.

Moreover, Keynes believes that fiscal policy is more effective than monetary policy during economic recession, because monetary policy may be failure due to the so-called "liquidity trap". The real cause of liquidity trap has been explained in Lecture 19, which is different from Keynes's idea.

With the support of seemingly reasonable Keynes's theories, governments of various countries have rightfully and crazily expanded expenditures, claiming that they have done so for the good of economy. As a result, the fiscal deficits of many countries rose sharply and accounted for a higher and higher proportion of GDP. For example, in Greece that has been in sovereign debt crisis, the fiscal deficit accounted for 160% of GDP in 2012. In other words, all in Greece have worked hard for one year and the wealth created is not enough to repay the country's debt. If it happened on an individual or a company, it is insolvent and should be bankrupt.

If Keynes resurrected and saw that there were full of countries overwhelmed by fiscal deficits and debt crises in the world nowadays, he might be very surprised why the governments did not increase tax and reduce expenditure when the economy is prosperous to have fiscal surplus covering the fiscal deficit in economic recession. It is naive to think that government really believes Keynes. In fact, government does not believe Keynes, but just makes use of him! Government is willing to follow Keynes's first half statement (when there is economic recession, government should increase expenditure without worrying about fiscal deficit), because it is convenient for it to get rid of tax restriction to spend freely. As for the latter half of Keynes's statement (when there is economic prosperity, government should increase tax and reduce expenditure to have fiscal surplus to cover the fiscal deficit before, and keep budget balance within an economic cycle), it will not care.

Even for an ordinary individual, when he gets used to spending lavishly, it is very difficult to suddenly cut expenditure and go back to live a prudent life. It is even more difficult for government that spends taxpayers' money and can easily continue to issue treasury bonds to deal with the fiscal deficit temporarily to cut expenditure and increase tax. Expenditure cut often implies cutting social welfare. Increasing tax and cutting social welfare are both objectionable to the public, especially in those "democratic" countries, so are difficult to be permitted in parliament. In fact, in Greece and even France, after the outbreak of the European debt crisis, the parliamentarians and political parties advocating austerity policy were simply voted out of office.

Section 6: The Crowding-out Effect of Government Expenditure

Even if government really believes in Keynes and strictly follows his recommendations to increase tax and reduce expenditure during economic prosperity to have fiscal surplus covering the fiscal deficit during economic recession, according to Friedman's permanent-income hypothesis, it means that government expenditure is a temporary expenditure that has nothing to do with permanent income and consumption and may not increase current income and consumption. Thus, the link of multiplier effect is interrupted.

For example, Keynes has advocated that during the Great Depression, it was good for government to employ labor to bury and then dig out bottles repeatedly, which could boost national income. However, when people expected government to employ them only temporarily and then they would be unemployed again, they would not dare to increase consumption, but would save the (temporary) income against the risk of future unemployment. As has been analyzed in Section 2, it is actually not saving but consumption on insurance that is not helpful to increasing future output.

Unless government established a state-owned enterprise to employ them, so that they would expect to work in it in the future and regard it as permanent income, then they would increase permanent consumption.

However, if government did so, the society would have become a planned economy. In the era of planned economy, there did not exist economic cycle in China, but there was permanent recession causing poverty and the national economy to be on the verge of collapse, which is the result of a huge increase in transaction cost caused by the infringement on PPRs. It proves that the rise in the proportion of transaction cost to social wealth is precisely the cause of economic recession.

Besides, many economists who criticized Keynes's theory of multiplier have pointed out that the increase in government expenditure would have "Crowding-out Effect" on private consumption and investment expenditures. If government spends on purchasing consumption goods, there will be less supply in the product market causing a rise in price that hit private consumption. As a result, consumption expenditure will decrease, which will also lead to several times of decrease in national income due to multiplier effect. If government spends on public investment, there will be less supply in the factor market causing a rise in factor price (especially the interest rate) that hit private investment. As a result, investment expenditure will decrease, which will also lead to several times of decrease in national income due to multiplier effect.

However, Keynes would argue that the prerequisite of his theory of national income determination is that there is economic recession, which means many resources are idle, so supply can increase at constant price. The question is: what is the state of "in idle"? It is logically similar to the question "What is unemployment?" In Keynes's view, for the unemployed, his labor is idle. If government hired the unemployed to bury and then dig out bottles, the enterprises would not have to increase wage to attract the labor, so there would be no increase in cost, neither would they decrease production and investment expenditure.

His argument seems plausible, but as the causes of unemployment has been explained in Lecture 20, readers should understand what mistakes Keynes has made. One is unemployed because his wage is too high, which means his contribution to production (the value he creates) is not enough to make up for his cost (wage). Government does not calculate cost and benefit as carefully as an enterprise because it spends taxpayers' money, and employs this labor with a high wage, which

cannot change the fact that the value created by him is less than the cost of his wage. In other words, the production that government employed this labor to engage in suffers a loss.

Similar analysis can be analogized to other idle production factors (resources). Although there are no evil polices such as minimum wage law imposed on most production factors, it is still possible that the prices of them cannot fall to the equilibrium price due to transaction cost, so that they are idle. If a resource is idle, the only direct reason is the cost of using it is higher than the benefit. The enterprises refuse to use the idle resource, but government does, which cannot change the fact that the value created by it is less than the cost of it, so it is definitely a business at a loss.

For an economy, it is by no means the more production, the better, but the more production with added value, the better. A production at a loss means consuming higher-value resources to produce lower-value products, which can only lead to negative growth of wealth. There do be some cases where the efficiency of government is higher than that of enterprises. However, since the relevant production have always been engaged in by enterprises, the efficiency of enterprises in such productions should be higher than that of government, otherwise the market economy should give way to the planned economy. The key question is: will the change in constraints of economic recession suddenly increase the efficiency of government in using these resources even higher than that of enterprises? Obviously NO!

It is logically similar to the consumption expenditure in the product market. If consumers are unwilling to buy a good, they must think that the use value of it is not worth its price. Can government change the fact that the use value of this good is not worth its price by buying it? Of course NO!

Thus, it is impossible that when there is economic recession government takes the place of individuals and enterprises to increase consumption and investment, which can increase economic growth rate. In fact, it just increases businesses at a loss that decrease social wealth.

The above is a complete criticism of Keynes's theory of national income determination.

Based on Keynes's theory, Keynesian school that is different from Keynes and is in fact not agreed by Keynes later developed IS-LM model. "IS" refers to the equilibrium condition of the product market ($I = S$), while "LM" refers to the equilibrium condition of the monetary market ($L = M$ or money demand = money supply). Keynes's theory of national income determination regards investment (I) as the independent decision of enterprise, so it is a constant. Keynesian school regards " I " as a function of interest rate (r), so interest rate will affect national income through investment. And interest rate is also affected by the monetary market, and national income will affect money demand, thus affecting interest rate. In order to solve the problem that national income and interest rate affect each other, Keynesian school came up with this IS-LM model to solve simultaneous equilibrium condition of the product market and the monetary market. Keynes refers fiscal policy to a tool to move the curve of IS, while monetary policy to move the curve of LM.

The equilibrium condition of $I = S$ has been thoroughly criticized in Section 2. As has been criticized in Lecture 19, if monetary policy is used to adjust the economy, it will only cause economic fluctuation (business cycle), so this IS-LM model is also completely wrong.

As for the analysis of aggregate demand and aggregate supply that is deduced from the theory of national income determination, of course it is also wrong, and it is unnecessary to waste time in introducing and learning it.

It is necessary to discuss the "Say's Law" that was an orthodox theory in economics before the Great Depression and the popularity of Keynes's theory. Say's Law says, "Supply will automatically create demand equal to it. An enterprise has paid to the upstream suppliers of raw materials or intermediate products and to the workers before production. According to the theory of income

distribution introduced in Lecture 20, the owners of these production factors will divide up all the income from the sale of the products. In other words, the purchasing power (demand) generated by the income of the owners of the production factors can definitely buy out all the products (supplies) in the market. Therefore, according to Say's law, it is impossible that there will be overproduction that aggregate demand is less than supply.

The GD seemed to have refuted this law that was once regarded as a golden rule in economics, so it is no longer introduced or even mentioned in economics textbooks nowadays, and few economics students have heard of it. However, according to the introductions and analyses of this lecture and the previous lectures about the GD and monetary theory, economic recession is not necessarily due to overproduction, but may be due to the rise in consumption that is not help to increasing future output. And even if there is overproduction, the real cause is wrong monetary policy.

There do be mistakes in Say's law, but it is just wrong in ignoring the non-neutral effect on the economy when the monetary value is unstable, which is the mistake of classical economics, and in ignoring the transaction cost, which is the mistake of neo-classical economics.

Lecture 22: Business Cycle and Economic Growth

Section 1: Financial Crisis cannot Destroy Real Wealth

In so-called macroeconomics, the short-term economic change is business cycle, while the long-term economic change is economic growth. However, both in fact refer to changes in national income or social wealth. If it increases, it is economic growth, while if it decreases, it is economic recession. There is no need to distinguish between short-term and long-term.

According to the discussions in Lecture 19 and Lecture 21, if there were no wrong monetary policy, there would be no economic fluctuation or business cycle. As for economic recession that refers to a decline in output, it is possible to be caused by a rise in risk (transaction cost) that makes people take defensive measures of buying insurance that is not helpful to increasing future output. The decline in output can also be regarded as an increase in transaction cost.

Here are some macro events in reality that further illustrate what key factors determine economic performance (whether short-term or long-term).

Example I: the Southeast Asian financial crisis in 1997

In 1997, international speculators led by Soros first attacked Thailand's currency, the Thai baht. Only fighting for a short time, the Thai government had given in and let the Thai baht depreciate sharply because of the exhaustion of foreign exchange reserves. The currency crisis in Thai baht quickly spread to other Southeast Asian countries that the export structures of them were similar to Thai. Their currencies were also attacked by international speculators one after another and forced to depreciate sharply. Even the Korean currency (won) was affected. In the end, the only national or regional currencies that survived this "catastrophe" were China's RMB and Hong Kong's dollar. This currency crisis also quickly turned into financial crisis, even economic crisis, and triggered political crisis in Indonesia.

During the crisis, the Southeast Asia countries suffered a great loss. The "East Asian miracle" was widely questioned. Mahathir, the Prime Minister of Malaysia had angrily accused Soros of sweeping away the economic wealth accumulated by Southeast Asian countries for decades overnight.

However, looking back this financial crisis in 1997 now, it can be concluded that the economic foundation of Southeast Asian has not been destroyed. Supported by the financial rescue from the International Monetary Fund (IMF), the Southeast Asian countries adopted the fundamental solution of cutting expenditure, reducing deficit instead of Keynes's prescription. What is more, the export competitiveness of them has greatly increased due to the currency depreciations, so their exports soared. The increase in trade surplus not only made up for the original deficit in balance-of-payment, but also brought foreign exchange reserves back to the pre-crisis level quickly. As a result, the economies achieved a strong V-shaped rebound. The doubts about the "East Asian miracle" vanished and were not mentioned again. Compared with South America, East Asia still has experienced rapid economic growth.

It is agreed that the financial crisis in Southeast Asia of 1997 seemed to be menacing and catastrophic at that time, but quieted down quickly due to the correct policies. From the perspective of long history, it was only a small episode on the long journey of Southeast Asian economy, and did not change the general direction of economic growth in the region at all.

Why? The answer is simple. As has been pointed out in Lecture 19, the real wealth is from goods (including tangible products and intangible services), rather than money that measures the value of wealth and facilitates transactions. Financial crisis does destroy something that is the number on the accounting books and used to represent the value of real wealth. And the real wealth of buildings, machinery and equipments and skilled labor (the production capacity that can create more new wealth) does not disappear, but just the digital value on the accounting books has declined (depreciation). As long as the financial crisis has passed away and the financial system (fictitious economy) based on the real wealth (real economy) stabilizes in a new equilibrium, all kinds of wealth whose physical form is completely intact (only devalued) can exert the production capacity as usual and create more new wealth again. Moreover, because they become much cheaper (due to devaluation), there is much less production cost, so the products are much more competitive internationally than before.

Conclusion: The crisis of fictitious economy (financial system) actually cannot destroy the physical form of real wealth, but only makes it devalued. As long as the chaos caused by the crisis is eliminated, it will not have a long-term and fundamental effect on economic growth.

Extension: The subprime mortgage crisis of US in 2008 that caused a financial tsunami sweeping the whole world and the European sovereign debt crisis in 2009 were both originally financial crises that would not destroy the real wealth. However, because the defects of the financial or monetary system that caused the crisis have not really disappeared, the fictitious economy is still out of line with the real economy, and the financial storm has never really gone.

More extensions: Canada, which is geographically and economically close to the US, has not been seriously affected by the financial tsunami in 2008 because there are few "toxic assets" caused by subprime mortgages in the financial system of it. European countries such as Sweden, which are also geographically and economically close to the Eurozone, have not been seriously affected because they do not share the same currency of Euro. It is proved from the opposite side that as long as the fictitious economy is not out of line with the real economy, the financial crisis from abroad will only have short-term shock, instead of long-term and far-reaching impact on a country's real economy.

In other words, business cycle is caused by the disconnection between the fictitious economy and the real economy, so government should try to make the virtual economy in line with the real economy.

Section 2: The Importance of Expectation/Confidence

Example II: the earthquake of China in 2008

Unlike financial crisis in virtual economy discussed in Section 1, the natural disasters such as earthquake actually destroy real wealth in physical form. However, it is also agreed that the earthquake of China in 2008, although destructive, there have been massive reconstructions there, and the economic growth has not even stopped. Similarly, from the perspective of a long history, it was only a small episode on the long journey of the local economy, and did not change the general direction of economic growth in the region at all.

Why? The answer is also simple. The earthquake did have destroyed some real wealth, but the remaining wealth still had the productive capacity to create more new wealth. As long as people are confident in the future and work hard, the destroyed wealth can be fully made up sooner or later. The key is people have positive expectations for future economic growth, and engage in creating more new wealth.

Conclusion: Although natural disasters such as earthquake do destroy real wealth, as long as people's confidence in future economic growth is not destroyed, the remaining real wealth can quickly create more new wealth to make up for the destroyed part.

Extension: The World War II was not a natural disaster that destroyed some (in fact, most) real wealth, but Japan and Germany as defeated countries that suffered the most were the fastest-recovering and rapidly rising economies after the war. Although most real wealth was destroyed in the war, both countries have established a more reasonable economic system (especially compared with Britain, where the leftist Labor Party came into power and implemented nationalization and government intervention), so people had positive confidence in the future and were willing to work hard. The houses and machines were almost destroyed, while skilled labor, as the most important production factor, were relatively well preserved and of high quality. As a result, these two countries took the lead in recovering, experienced rapid economic growth and caught up with the US quickly.

One more extension: The devastating earthquake of Japan in 2011, together with the following tsunami affected millions of thousands people. Ten times of the damages of financial crisis in 1997 might not match that of this natural disaster. The following nuclear leakage was even much worse, because it was continuously threatening to destroy real wealth (including real goods and human lives), which could strike people's confidence in the future more than a one-off catastrophe. What is more, Japan has been in economic depression for 20 years since the 1990s, and people's confidence in economic growth has already been hit hard. With pessimistic expectations, the production factors such as capital and high-quality talents that were easy to flow (easy to run away) would flow out of Japan in large quantities.

As a result, compared with the earthquake of China in 2008, that of Japan in 2011 was quite different in that it not only has destroyed real wealth, but also destroyed people's confidence.

Section 3: The Determinants of Economic Performance

The examples in Section 1 and Section 2 show that there are the determinants of economic performance:

(1) The productive capacity of wealth in the real economy

According to the theory of interest, real wealth is the sum of the discounted value of future income, and asset is anything that can bring future income because it has productive capacity. Even if the value of wealth is depreciated, as long as there is productive capacity, the negative effect on economic growth may only be temporary.

The traditional economic growth theories attribute the source of sustained economic growth to technological progress, because increasing investment in capital or labor can just lead to economic growth for a while. According to the law of diminishing marginal product/return, the increase in the input of any production factors will finally lead to the economic growth slowing down. By contrast, technological progress can break through the constraint of this iron law by increasing the productivity of wealth. The increase in the input of capital or labor can just move the point along the same curve of marginal product, while technological progress can move the whole curve to the right.

By the way, Keynes put forward the law of diminishing marginal propensity to consume to explain the insufficiency of aggregate demand caused by insufficient consumption, and put forward the law of diminishing marginal efficiency of capital to explain that caused by insufficient investment. The latter law is actually an implication of the law of diminishing marginal return, but Keynes completely forgot the role of technological progress, without which breaking through the constraint of this law, the human society should have been downgraded to as poor as animals'.

(2) People's expectation/confidence

Wealth is not living thing, so it depends on human to make use of its productive capacity. If people are pessimistic about the future, as mentioned in Lecture 21, they will take defensive measures of buying insurance that is not helpful to increasing future output. What is more, the assets (wealth) that are easy to flow will flow out of the country, thus directly reducing the wealth within a country.

(3) Whether the fictitious economy is out of line with the real economy.

As mentioned in Lecture 19, the fictitious economy is created by human to facilitate the real economy, or more clearly to reduce the transaction cost of the real economy. However, the value of the fictitious economy is derived from the real economy, so if the fictitious economy is out of line with the real economy, people will be misled in decision-making based on the real economy, which means there is increase in information cost, so people are unable to make the best use of the productive capacity of wealth, resulting in a decline in output or poor economic performance.

The cause of the Southeast Asian financial crisis in 1997 was that the exchange rates between the currencies of the Southeast Asian countries and the US dollar were overvalued before the crisis, which was out of line with their economic strength. After the currencies depreciated to be in line with their economic strength, the governments did not increase expenditure as Keynes' theory suggested, so the economy recovered quickly.

The cause of the financial turmoil of the US in 2008 was that the high-risk subprime mortgage and the financial innovation of asset securitization led to excessive issuance of “toxic assets” that was out of line with the real economy of the real estate market. As long as there are still toxic assets in financial system, the financial market is out of line with the real estate market, and the economy cannot recover.

The cause of the European sovereign debt crisis since 2009 was that the exchange rate of euro was too high for the “PIIGS” with weak competitiveness, which means the exchange rate of euro was out of line with the real economy of PIIGS, similar to the financial Southeast Asian crisis in 1997. However, the Southeast Asian countries have their own currencies, so their economy can quickly recover by depreciating their currencies to be in line with their economic strength. By contrast, PIIGS do not have their own currencies, so the only way to solve this problem is to strengthen their economy to be in line with the strong euro.

For example, suppose the average price of the Greek products is 2 euro, and the exchange rate between the US dollar and Euro is 1:1, so the average price of the Greek products is 2 dollars in the international market. And suppose the average price of Chinese products comparable with the Greek products is 1 dollar, it is obvious that Greek products cannot compete with Chinese ones. If Greece had its own currency, it could depreciate its currency to the exchange rate of 1:2, which meant the average price of the Greek products would become 1 dollar in the international market now, so they could compete with the Chinese ones. But as a member of Eurozone, Greece has no right to depreciate Euro. What should it do? It can just try to reduce the average price of the products to 1 euro, so that even if the exchange rate remains 1:1, the average price of Greek products in the international market can fall to 1 dollar, and they can compete with the Chinese ones. How can the average price of Greek products be reduced to 1 euro? It is necessary to cut costs, including lowering the price (wage) of the labor force, so it is necessary to abolish the minimum wage law, the trade unions with administrative monopoly, and cut social welfare to reduce the cost burden on enterprises. Even if social welfare is borne by government, the cost is actually paid by taxpayers including enterprises and individuals. Obviously, it needs a painful and lengthy major reform of the institutional and economic structure, which cannot be achieved overnight.

(4) The proportion of transaction cost to national income or social wealth.

In fact, the determinants of (2) and (3) are also related to transaction cost. People are pessimistic about future, which implies the risk (transaction costs) rise. The fictitious economy is out of line with the real economy, which means the function of the fictitious economy facilitating the real economy does not well, or there is an increase in institutional cost (transaction cost). Combined with the determinant of (1), it is easy to understand that there is objectively the productive capacity of wealth, but human may not be able to fully make use of it due to the constraints of technique (the factor in nature science) and transaction cost (the factor in social science).

However, the effect of transaction cost on making use of the productive capacity of wealth is not limited to the determinants of (2) and (3). In reality, if the transaction cost is so high to affect the economic performance, the most common case is the ineffective protection of PPRs. There is an example of one's knowing how to grow grain on stones in Lecture 17, which has well demonstrated how the transaction cost caused by the lack of protection of PPRs prevents people from making use of the productive capacity of wealth (including the technology and the stones).

After all, it is only a fictitious example. Here is another example in reality. When China has just reformed and opened up, there have been some Hong Kong-invested enterprises that set up factories in the Pearl River Delta region. At the beginning, the Chinese government did not allow them to recruit workers freely. Instead, the government sent the “state employees” to work in the factories. The working hours and wages of the state employees were decided by government, and the Hong Kong investors cannot fire them. Therefore, the state employees worked loosely and

took a long nap at noon. Finally, the Hong Kong-invested enterprises refused to hire state employees and strived to hire "contract workers" by piece-wage contract. The situation changed greatly immediately: the factory did not open until 8:00, but the workers were already waiting outside at 7:00. Not only did they never take a nap at noon, but also they wanted to work overtime when it was time to get off work. The output increased by more than double. In the mid-1980s, the economic growth rate of many regions in Guangdong was over 50% every year! The institution of employment changed from "state employees" to "contract workers", which implies the PPRs of labor have been clearly delineated, so the transaction cost has been greatly reduced, and both the productivity of wealth (labor) and economic growth rate have rapidly risen.

Ineffective protection of PPRs will lead to an increase in transaction cost, which is actually due to rent dissipation. The relevant analysis will be introduced in detail in the later lectures.

As mentioned in Lecture 21, the "permanent income" in Friedman's "permanent-income hypothesis" is actually the "annuity" introduced in Lecture 18. If the economic performance is reflected by wealth changes, Fisher's wealth equation $W = \frac{Y}{r}$ can be adopted to analyze this issue.

The expansionary fiscal policy advocated by Keynes is to increase national income (Y) several times by increasing government expenditure, which has been criticized that government expenditure can only increase temporary income, neither can affect temporary consumption nor permanent income (annuity), so cannot increase permanent consumption.

How about expansionary monetary policy? According to Fisher's wealth equation, when income (Y) falls sharply due to the outbreak of economic crisis, if interest rate (r) does not fall or falls even further, wealth (W) will fall sharply. Therefore, the logic of expansionary monetary policy is to increase money supply to make interest rate (the price in monetary market or capital market) fall at least as much as income, so as to keep wealth unchanged, thus ensuring that consumption and investment expenditures will not decrease. However, if monetary policy is used to influence interest rate instead of stabilizing monetary value, it will cause the fictitious economy out of line with the real economy sooner or later, which will bring more adverse consequences. What is more, as has been pointed out in Lecture 19, if people are pessimistic about the future, the effective money supply in the economy may not increase due to expansionary monetary policy.

The key to boosting the economy is to change people's expectation of future income, the essence of which is to increase permanent income. The general principle is to reduce transaction cost, one of the results of which is expectation change. What government can change most effectively is the expectation of enterprises, because by abolishing the minimum wage law and other evil laws that interfere with the operation of enterprises, enterprises will reverse their pessimistic expectations for the future, thus expanding production and increasing supply.

According to the theory of national income determination, all the Keynes' policy suggestions focus on increasing aggregate demand. However, according to the Say's Law, increasing supply can automatically create demand, as long as there is not wrong monetary policy so that money remains neutral. Of course, it is not government that increases supply, but enterprises that have changed the pessimistic expectations, which can stimulate effective investment expenditure. In order to produce, enterprises must first purchase all kinds of production factors, which immediately increases employment and the incomes of the owners of all kinds of factors (including workers who are the owners of labor), so that the owners can increase consumption expenditure. The incomes obtained by the owners of these factors are permanent income, because as long as enterprises keep producing, they will keep purchasing production factors and providing income to the owners. By contrast, there may be government's investment this year, but there may not be the next year. If government keeps investing, it will replace enterprises' production, which means the planned economy will replace the market economy.

In fact, there is an anti-Keynes school called "Supply-side School", and the essence of which is precisely the idea of Say's Law.

Lecture 23: Coase Theorem

From this lecture on, the paradigm has changed from the traditional economics to the new institutional economics. Traditional economics focuses on resource use and income distribution, but lacks the important part of institutional arrangement, which is due to ignoring transaction cost. As mentioned in Lecture 14, transaction cost in broad sense is institutional cost. In fact, even in the previous lectures, we have frequently talked about transaction cost, so it is impossible to completely separate the new institutional economics from traditional economics, and just the contents that are generally classified as new institutional economics will be discussed from this lecture on.

What is more, in the “modern” economics textbooks, there will be some contents involving new institutional economics in the so-called microeconomics, which are classified as “welfare economics”. The mistakes of welfare economics have been criticized in Lecture 5, and will be criticized in detail in the later lectures.

Furthermore, the name of new institutional economics comes from that fact that it concerns about institutional arrangement, and more specifically, institution is contract, so institutional arrangement is the same as contract choice. Thus, in broad sense, “contract economics” is the same as new institutional economics. Similarly, transaction cost is institutional cost in a broad sense, so transaction cost economics is the same as new institutional economics. There is also a “property rights economics” that focuses on the delineation or protection of PPRs. When PPRs are clearly delineated and effectively protected, the price criterion can be used to determine the competition, and there will be no transaction cost caused by rent dissipation. Of course, there is transaction cost in the delineation or protection of PPRs. Thus, property rights economics is a subset of new institutional economics or contract economics or transaction cost economics.

Section 1: The First Version of Coase Theorem

The new institutional economics is based on the concept of transaction cost, and an important theory based on the concept of transaction cost is Coase Theorem.

There are actually three versions of Coase Theorem, and in the economics textbooks, there is usually only the third version. In this book, three versions will all be introduced one by one.

(1) Version I: The delineation of right is an essential prelude to market transactions.

This version was put forward by Coase in the paper "The Federal Communications Commission".

When the new institutional economics was introduced into China, there was a debate in economics circles: Which came first, the delineation of PPRs or the market transactions? It seems to be a question of the chicken or the egg. From the perspective of time order, it may not be easy to judge, but from the perspective of logical order, the delineation of PPRs must come first, then there are market transactions. If one had no right to a good (whether it is clearly delineated by law or by the tacit approval of enough people), why will the others buy it from him? The others who need the good can just rob him instead of paying. The law of the jungle is the competition criterion of Nature, and there is no institution needed as the game rule to support the establishment of this criterion. By contrast, the other competition criteria (not necessarily price criterion) need certain institutions as game rules. For example, the criterion of first-come-first-served implies that everyone tacitly approves those come first have priority.

According to the introduction of competition criterion and game rule in Lecture 4, it is obvious that this first version of Coase theorem is actually a specific example that game rules support the establishment of competition criterion. Here game rule refers specifically to the delineation of right, and competition criterion refers specifically to price criterion (market transaction). Only with the game rule of delineation of right, can the price criterion of market transactions be effectively established.

When right is delineated, there are PPRs. On the other hand, all goods, assets or resources transacted (transferred) in the market have PPRs. Whether government has approved with law or not is another matter.

PPRs are composed of three rights: use right, right to income and right to transfer.

Use right refers to the right of the owner to decide the resource use, including (i) whether to use or not (the owner can decide to leave the good in idle); (ii) who to use (the good can be used by the owner or the others or even the public); (iii) how to use.

Attention: competitive behaviors need to be restricted. In the previous lectures, the term "protection of PPRs" is usually used, which is actually one-sided. The all-sided term should be "delineation of PPRs". Protection is only one side of delineation, and the other side is restriction. Coase chose the word "delineation" very accurately, which means that law draws a line, and on one side of this line is what the owner can do, and law will protect him, while on the other side of the line is what the owner cannot do, and law will restrict him. Where this line should be drawn is the most test of the wisdom of law makers. Too much restriction is infringement on the owner's PPRs, but too much protection is also infringement on others' PPRs.

For example, one has an apple, which means he is the owner of the apple. He can decide to eat it, sell it, give it to others, or even throw it away. However, he cannot use the apple to hit another's head, and neither can he throw it as trash in public places. If one has complete and absolute PPRs to the apple, he should do anything with it as he wishes. So why can he not hit another with the apple? Because what he does has infringed the PPRs of another's body and even the life. And the behavior of throwing the apple as trash in public places has infringed the rights of the public to have clean public places. Therefore, on the one hand, laws protect PPRs, but on the other hand, they also restrict them mainly by regulating how to use a property. The delineation of use right is also the restriction of it, including the kinds of use (e.g. one is not allowed to use his apple as a weapon to attack others) and the limit of each kind of use (e.g. one is allowed to throw away his apple, but is not allowed to throw it as trash in public places).

In reality, there are many cases not as simple and clear as the above example of apple. For example, how to draw a line for pornography? In different countries, there are very different legal provisions. In China, it is illegal to disseminate pornography in public as a means of profit. In the US, Hong Kong and many countries or regions, there is a grading system, where it is allowed to disseminate pornography to adults in public as a means of profit, but it is forbidden to do so to juveniles. In Japan, the "excellent" pornography can be even awarded prizes.

According to equal marginal principle, the line should be drawn where the social marginal benefit brought by protection is equal to the social marginal cost caused by restriction, so that the social benefits can be maximized. Based on the specific constraints, the above tautology can be turned to refutable, so scientific explanation.

For example, in China, with so strict prohibition on pornography, many artistic works (such as western nude art and some classic works in ancient China) are mistaken as pornography and were once regarded as "obscene" works and listed as banned books, which have suppressed the development of this kind of art that is an important non-monetary income for a society.

As a sophist in ancient China said, "The desire for food and sex is part of human nature." The industries such as prostitution have stubbornly survived since ancient times, and always revive again and again after strict bans that cannot have long-term effect. Before the reform and opening up in China, prostitution was completely banned for quite a long time. At that time, compared with the very good social security in other fields, the crime rate of rape was particularly high, because men with such need could not find legal way and had to put their hands on good women. Such laws against human nature are difficult to effectively implemented, and the social cost is too high.

It should be repeated that where the line should be drawn is the most test of the wisdom of law makers, not moralists.

Right to income is the core of PPRs, because the purpose of use right and the right to transfer is to ensure right to income. An asset is valuable because it can bring income. One has PPRs of this asset because he can obtain income from it. Even if right to income is only partially infringed (such as price control that has infringed part of income where the regulated price is lower than equilibrium), there will be non-private or public effect on PPRs. The infringed part of income becomes ownerless, which causes self-interested people to compete for it, resulting in rent dissipation that implies the rising transaction cost.

A question: does government's tax infringe right to income right? In ancient times, "rent" and "tax" were often synonymous, because the earliest governments were often also landlords. As landlords, they provided lands as production factors and collect rents, while as governments, they provided public services such as public security and dispute resolution and obtain corresponding incomes. Moreover, tax clearly belongs to government, so there is no ownerless income that will

lead to rent dissipation. Of course, government is a virtual concept, and it is government officials that really exist, so the delineation of PPRs of tax revenue is somewhat ambiguous.

If government collects specific tax that levied according to quantity such as the population or the area of lands, and the value of public services it provides is worth what taxpayers pay, tax is only the price for public services, and does not infringe right to income. Of course, it is possible for government to collect unnecessary or exorbitant tax out of line with the value of public services. If people could easily run to other countries, this behavior of voting with feet can effectively restrict government. However, it is not easy to do so in reality, which implies there is high transaction cost.

If government collects ad valorem tax that levied according to price and is actually commission based on income or share revenue, since government is actually not the owner of the resources that bring income and has no right or is difficult to decide the resource use, this kind of tax does infringe right to income.¹

Right to transfer refers to the right of the owner to transfer use right to others for a period of time (lease), or to transfer all the three rights of PPRs thoroughly to others (sale). Right to transfer can directly reflect the essence that market transactions determine the outcome of competition by price criterion. Because the outcome is determined by price, resources can fall into the hands of those who are best at using them. Whoever can make full use of the productive capacity of resources can obtain the highest income from the resource use, so he can offer the highest price to defeat the other rivals.

The economic implication of the first version of Coase Theorem is that all goods (assets and resources) transacted (transferred) in the market have PPRs. In other words, all goods with right to transfer must have use right and right to income to some extent. In fact, no one is willing to buy something that has no use and brings no income.

However, it may not be true that all goods with use right must have right to transfer. For example, the private office of a teacher in university has use right and right to income, which means the teacher can use the office and enjoy the relevant non-monetary income, but he or she cannot transfer the office. The PPRs without right to transfer is incomplete, and the value will be lower, because the resource cannot be transferred to those who are best at using it, and the productive capacity of it cannot be fully used to maximize the value (wealth).

One more question: does leasing and selling are different for the value of an asset? If the tenant can sublet the asset, and the lease can be renewed after expiration, there is no difference in theory. For example, renting a house permanently is essentially the same as selling a house for once. However, it is necessary to negotiate for subletting and renewing the lease after expiration, so the transaction cost of leasing may be higher than that of selling that needs to negotiate for once. If the lease period is relatively long, or if the provisions for renewal are so clear that the lease can be automatically renewed without negotiation, the transaction cost can be low enough, and there is little difference.

It is an important implication that ownership does not matter!

Ownership is only a nominal right in law. As long as there are three rights (use right, right to income and right to transfer), even if there is no nominal ownership, it does not matter. The most typical example is that the ownership of lands in China and Hong Kong is not private in law. In China, lands are state-owned in law, while lands in Hong Kong belonged to the British King or Queen before the return to China and now belong to Hong Kong Special Administrative Region Government. Thus, the ownership of lands belongs to the state or the government, but the state or the government can auction the lands (in fact, rent them out). As long as the private have use right, right to income and right to transfer the lands, the lands have PPRs. The term for most of

lands in China is 70 years that can be extended according to the "Property Law", so it is the same as the case that subletting is allowed (both the land and the houses on it are allowed to be transferred), and the lease period is long and the lease can be renewed after expiration.

1. In his book "The Economic System of China", Steven. N. S. Cheung has made an in-depth study of value-added tax in China. He pointed out that the essence of it was a series of sharing contracts between the local governments and investors (businessmen), and among governments at all levels. The land contracts signed by local governments and investors showed that local governments had taken part in the decisions about the land use and the businesses, so value-added tax did not infringe right to income. It is a very complicated issue that is not suitable to be introduced in this entry-level economics textbook. Those who are interested in this issue are recommended to read Cheung's book. [←](#)

Section 2: The Second Version of Coase Theorem

(2) Version II: With clear delineation of rights, the resource use through market transactions will bring the highest value.

This version was also put forward by Coase in the paper "Federal Communication Commission".

However, it is actually a tautology, because with the postulate of self-interest, all behaviors (including the behavior of resource use through market transactions) must be optimal under the current constraints. Therefore, under the constraint of "clear delineation of rights", resource use must bring the highest value, otherwise it is not optimal, which will be out of line with the postulate of self-interest. Thus, this theorem is a tautology, and cannot be wrong or refutable.

On the other hand, since the postulate of self-interest has determined all behaviors must be optima under the current constraints, even without clear delineation of rights, and the resource does not used through market transactions, it will still bring the highest value. Even if there is rent dissipation due to lack of delineation of rights, and the value of resource or asset drops to zero, it is still the highest! Compared with negative, zero is the highest. It is just another constraint that there is no delineation of rights.

Of course, Coase might mean the value of resource without clear delineation of rights will be lower than that with clear delineation of rights. However, there is transaction cost for the clear delineation of rights, so if it is higher than the transaction cost (rent dissipation) in the case without clear delineation of rights, the net value of resource or asset may not be the highest.

It is necessary to introduce an important concept of "Pareto-optimality" in welfare economics. In welfare economics, efficiency is defined as whether it is Pareto-optimal. As mentioned in Lecture 5, welfare economics is wrong in judging "to be or not to be" with efficiency, because it is impossible that there is inefficiency or waste with the postulate of self-interest. This mistake will be discussed in detail here.

When introducing the concept of "optimum" in Lecture 7, it has been explained that there is no difference between optimum and self-interest. How can a self-interested person choose a non-optimal option? Thus, individual optimum is the same as the postulate of self-interest.

How about social optimum? The Pareto-optimality here precisely refers to social optimum. State A in a society is "Pareto-preferred" to some other state B if, under A, everyone is at least as well off as under B and at least one person is better off. Thus, a state is "Pareto-efficient" or "Pareto-optimal" if no available alternative is Pareto-preferred to it. In other words, everyone in the society has been optimal without at the cost of harming the others. If one can make a change that at least makes another one better off while all the others as well off as before, it means the society is still not in the state of Pareto-optimality, and he should make that change ("Pareto-improvement").

The postulate of self-interest ensures that all behaviors must be optimal for individuals, but can it ensure that they must be optimal for the society? With transaction cost taken into account, the answer is yes. If a change benefits someone without harming the others, the self-interested beneficiary must make it happen and no one will stop him. The problematic case is: if a change benefits someone but will harm another, but the beneficiaries gain a lot, while the harmed only suffer little, so the benefits are more than harm. If the beneficiaries take part of the benefits to completely compensate the harmed, this case is still Pareto-optimal, so the change should happen. However, in reality, it seems that this kind of change sometimes does not happen.

Why? The reason must be that there is too high transaction cost of negotiating compensation between the beneficiaries and the harmed. How much benefit and harm does the change cause? There is high information cost to know it, and negotiation will also cause high transaction cost. If the transaction cost of reaching an agreement is higher than the net benefit after the beneficiaries completely compensate the harmed, the society is actually in a worse state due to the change. Therefore, the state after the change seems to be better off than that before the change only because the transaction cost of the change has not been taken into account. In other words, the constraints before the change do not include the transaction cost of the change, while the constraints after the change do. Different constraints are not comparable.

Based on the above analysis, it is easy to understand the problem of the second version of Coase Theorem. The value with clearly delineation of rights and resource use through market transactions is higher than that without them, and the latter seems to be inefficient or wasteful. However, there is transaction cost of delineation of rights and market transactions. If it is too high (higher than the increase in value with them), it is not worthy. The constraint with clear delineation of rights is different from that without it. More specifically, the former constraint is that "the transaction cost of delineation of rights is low, so rights are clearly delineated", while the latter constraint is that "the transaction cost of delineation of rights is high, so rights are not clearly delineated".

In a word, as long as all constraints are taken into account (especially the constraint of transaction cost), it must be Pareto-optimal, which means there can be no so-called "inefficiency" or "waste", so the value must have been the highest under the current constraints.

What is more, as mentioned before, the equilibrium condition is equivalent to the equal marginal principle. Since equilibrium is the same as optimality, Pareto-optimality is also equivalent to the equal marginal principle. Welfare economics introduced the concept of Pareto-optimality in order to prove that the market structure of pure competition is Pareto-optimal because the marginal cost (MC) of producers in price-taking is equal to the marginal use value (MR) of consumers. The economists of MSE think if there is no market, or the market transactions fail to satisfy the equal marginal principle (such as monopoly), these cases are not Pareto-optimal, and are criticized as "Market Failure". They advocate that government should interfere with the market to make it Pareto-optimal.

However, according to the analysis in this section, there will always be Pareto-optimality if all constraints (including transaction cost) are taken into account. Some market transactions do not seem to satisfy the equal marginal principle, because either the economists of MSE they misunderstand the modes of market transactions¹, or they neglect the transaction cost. In some cases where there seems to be no market, maybe there do be markets, but the transaction cost of direct transaction is too high and indirect transaction is used, or the transaction cost of market is too high and self-interested people choose not to use the market. The latter case has nothing to do with market failure, just as one chooses not to buy a good if it is too expensive, which is not the case that it is "good failure".²

1. For example, as has explained in Lecture 16, in reality, monopolists will adopt use various pricing arrangements to eliminate deadweight loss, so that the MC of producers is actually equal to the MR of consumers. ↩

2. Attention: Keynes believes that the free market will cause aggregate demand less than aggregate supply, thus causing overproduction and economic recession, which is totally different from the "market failure" criticized by the welfare economics. Keynes's mistakes have been analyzed thoroughly in Lecture 21. ↩

Section 3: The Third Version of Coase Theorem

(3) Version III: If property rights are well-delineated, and if the parties involved can reach and enforce agreements at zero transaction costs, then the final outcome will be efficient regardless of the initial assignments of property rights.

This version of Coase theorem is the most well-known and generally introduced in economic textbooks. It was put forward by Coase in his paper "The Problem of Social Cost". It is also known as "Invariance Theorem", because regardless of the initial assignments of PPRs, the final outcome of market transactions will be invariant in that it must satisfy the equal marginal principle or be Pareto-optimal.

Here are some examples put forward by Coase to illustrate this version of Coase Theorem.

A piece of land can be used to grow wheat or for parking. If used for parking, the wheat will be crushed by the cars. And if used to grow wheat, there will be no place for parking. Most will think it is the owner of car who has harmed the interest of the owner of wheat, but it is just because most think farmers are vulnerable, so put themselves into farmers' shoes, which is not objective. The objective analysis should be from the perspective of the whole society. If the rent of growing wheat (the income from selling wheat in the market) is higher than that of a parking lot (the income from parking fee), the land should be used to grow wheat, and vice versa.

According to this version of Coase Theorem, as long as the rights of this land is clearly delineated, no matter whether it belongs to the owner of wheat or that of car, the final outcome of market transactions will be that the land is used for the purpose with the highest rent.

Suppose the rent of a parking lot is higher. If the property rights of the land belong to "A" who is the owner of car, of course he will use it directly as a parking lot. But what if the property rights of the land belongs to "B" who is the owner of wheat? "A" can pay to rent or buy the land from "B" and turn it into a parking lot. Will "B" transfer the land to "A"? As long as "A" pays him more than the income from growing wheat, "B" has no reason to refuse. So will "A" pay? Since the income (rent) of the land as a parking lot is higher than that of growing wheat, "A" can still have net income after fully compensating the loss of "B" without the land used to growing wheat, and he has no reason not to pay. Thus, as long as the rights of the land are clearly delineated, no matter whom they are originally assigned to, through market transactions the land must be used for the purpose with the highest rent, which will not harm anyone, so is Pareto-optimal.

Extended from the above example, there is another example. There are two adjacent pieces of land, one is used to grow wheat, and the other is used as grassland to raise cattle. The cattle on grassland often go to eat the wheat. Most may think that "A" who is the owner of cattle has harmed the interest of "B" who is the owner of wheat. However, it can be also regarded that "B" has harmed the interest of "A" by infringing on the right of the cattle to eat wheat.

If the rights of wheat belong to "B", which means the cattle cannot eat wheat without B's permission. The cattle can get fatter if they can eat wheat, and "A" can get higher income by selling the cattle in the market. Suppose the increase in income (marginal revenue, MR) of "A" is greater than the decrease in income (marginal cost, MC) of "B" caused by the cattle eating wheat, "A" can pay to "B" in exchange for the permission of allowing the cattle to eat wheat. As long as the payment can make up for the loss (MC) of wheat damage, "B" has no reason to refuse. And as long as the payment is lower than the benefit (MR) of "A", he has no reason not to pay.

However, according to the law of diminishing marginal return, with the cattle eating more and more wheat, the MR of "A" is falling, and the MC of "B" is rising. The payment that "A" can agree is getting lower and lower, while that "B" will require is getting higher and higher. Eventually, when $MR = MC$, "A" will stop paying to get the permission from "B" of allowing his cattle eat more wheat. In other words, "B" will not rent or sell all the wheat in his land to "A" to raise cattle, but rent it until where the MC of wheat is equal to the MR of the cattle that get fatter by eating wheat, and they will build a railing to prevent the cattle from coming to eat more wheat.

Now let us reverse the assignment of the property rights, and suppose that the rights of wheat belong to "A", which means the cattle can eat wheat at will without B's permission. If cattle eat all the wheat in B's land, the MR of "A" may be lower than the MC of "B", so now it is "B" who needs to pay to "A" in exchange for restricting the cattle from eating wheat. Since the MR of "A" is lower than the MC of "B", they can reach agreement as long as the payment is higher than the MR of "A" and lower than the MC of "B". With the cattle eating less and less wheat, the MR of "A" is rising, and the MC of "B" is falling. The payment that "B" can agree is getting lower and lower, while that "A" will require is getting higher and higher. Eventually, when $MR = MC$, "B" will stop paying to "B" for restricting his cattle from eating wheat. Thus, they will build a railing in the same place as the above case that the rights of wheat belong to "B".

Both the above examples are fictional, and here is the third example in reality. Pollution from factory exhausts is harmful, and most think it is the factory that infringes on the residents living nearby. However, if the increase in output (MR) of the factory is greater than the increase in harming health of the residents (MC), from the perspective of the whole society, the factory should increase pollution (increase production). As long as the rights of air are clearly delineated, through market transactions there will be Pareto-optimality that satisfies the equal marginal principle.

In reality, the rights of air are usually delineated to the residents. If the MR of the factory of increasing production is higher than the MC of residents, the factory will pay to (compensate) the residents to buy the rights to pollute the air. As long as the payment (compensation) is higher than the MC of the residents and lower than the MR of the factory, there is no reason for both parties not to agree.

On the other hand, if the rights of air are delineated to the factory, which means it is free to pollute the air, and suppose the MR of the factory is lower than the MC of the residents, now it is the residents who need to pay to the factory in exchange for reducing production. Since MR is lower than MC, as long as the payment is higher than MR and lower than MC, there is no reason for both parties not to agree.

The last example is the real case that Coase discussed in his paper "The Federal Communications Commission". At the beginning of the 20th century, fishermen in Boston would contact with their families through radio wave when they went fishing. Many may use the same wave frequency, resulting in mutual interference. The US government then set up the Federal Communications Committee to regulate the use of radio waves.

When Coase studied this history, he found a file that recorded a parliamentarian's question, "Why not publicly auction the frequencies of radio wave, so that the taxpayers can get the benefits?" Government will publicly auction the lands, so why not do the same to the frequencies of radio wave? This question inspired Coase to find the essence of the problem. Different people use the same wave frequency and interfere with each other, because the rights of the frequencies are not clearly delineated. If the rights of frequency are clearly delineated just like the rights of lands, regardless the initial assignments of the rights, the frequencies can be transacted in the market and transferred to those who can have the highest income from the use because they can bid the highest price.

This example is important because it not only comes from the reality, but also can easily avoid the interference of subjective emotions compared with the previous three examples. In the previous three examples, it is easy to think one party is good, while the other is bad, and the bad party has infringed on the good party. Based on such moral judgment, naturally the policy suggestion to government will be to restrict or even punish the bad party. However, in the case of frequencies of radio wave, it is clear that all are the same in using the frequencies and interfere with each other, so the essence of the problem is obviously that the rights of frequencies are not clearly delineated, and the policy to solve the problem should focus on the constraint of the delineation of rights.

There is an important prerequisite of "zero transaction costs" for the third version of Coase Theorem. However, in reality, the transaction cost is not zero, which does not mean that Coase theorem will fail. The textbooks of MSE precisely make such a mistake in welfare economics. Firstly, they talk about the market failure, then introduce Coase theorem, pointing out that there will be no market failure if the Coase theorem is satisfied, and finally advocate once there is transaction cost, there will be Coase theorem failure.

In fact, one of the economic implications of Coase theorem is precisely that if the transaction cost is not zero, different initial assignments of property rights will matter, but the final outcome will be still Pareto-optimal or the equal marginal principle is still satisfied. Specifically for the second example above, where to build the railing will be different as the rights of wheat are assigned to "A" or "B".

The most typical example is the one of factory pollution. Why are the rights of air usually delineated to the residents nearby in reality? Not because residents are vulnerable and need protection, but because there is only one factory and there are many residents. The transaction cost of negotiating with one factory to pay to the residents will be lower than that negotiating with many residents to pay to the factory. In other words, in reality, people will choose to delineate the property rights with less transaction cost.

What is more, in reality, even if the rights of air are delineated, the transaction cost of negotiation is still too high, so there are few market transactions of transferring the rights of air, but there are usually judgments by court. The residents bring a lawsuit against the factory, and the court makes a judgment on the compensation the factory should pay to the residents. Coase's paper "The Problem of Social Cost" has carefully examined the courts' judgments on various infringement cases, and found that although the judges are not economists and certainly do not know the concept of transaction cost, but they do have made the judgments according to the consideration that to which party the rights are delineated is better for the society or cause less transaction cost.

As has mentioned in Lecture 14, the purpose of any institution (including market and court) is to reduce some kinds of transaction cost, but it will also cause another kinds of transaction cost to increase. Self-interested people will trade off the benefit (the decrease in transaction cost) and cost (the increase in transaction cost) of an institution, and decide whether to choose it or not. There is transaction cost of delineation of rights for market, while there is also transaction cost of judging arbitrarily for court. Unlike market transactions where there is the equal marginal principle to ensure the Pareto-optimality or social interests maximization, courts' judgments inevitably diverge from this principle more or less without the guidance of price in the market, which implies there is transaction cost. People will trade off the benefit and cost to decide to choose market or court.

In other words, if there were no transaction cost, there would be no different institutions or the effects (benefit and cost) of different institutions would be the same, so people do not need to make choices, and their behaviors will be random and unpredictable. Thus, there is a logical contradiction in the third version of Coase theorem: with zero transaction costs, there would be

no need for the institution of market, and neither is there need for a legal system to delineate the rights!

Steven N. S. Cheung has revised the third version Coase theorem as: If property rights are well-delineated, with the relative price of the market unchanged, then the final outcome of market transactions will be invariant and efficient, regardless of the initial assignments of property rights.

There is an additional prerequisite of "with the relative price of the market unchanged" for eliminating the wealth effect. Some quibblers argue that those to whom the property rights are delineated will become richer, and the use value of goods will change marginally under the wealth effect, which may cause the (relative) prices of these goods change in the market, resulting in different outcomes of market transactions.

Thus, the core content of the third version of Coase theorem can be preserved without the assumption of zero transaction costs, but only with the assumption of unchanged transaction cost or zero marginal transaction costs.

Lecture 24: Externality

Section 1: The Cause of Externality

As mentioned in the previous lecture, welfare economics likes to criticize market failure and provides a seemingly scientific basis for government to regulate (intervene with) the market. Monopoly is said to be one kind of market failure, which is wrong and been refuted in Lecture 16. There is another kind of market failure called "Externality".

The "external economy" or "external diseconomy" introduced in Lecture 15 is somewhat similar to externality, but not quite the same. Externalities arise when individuals or firms are involuntarily affected, either favorably or unfavorably, by the decisions of another party – where the decision-maker causing the externality is not penalized for the damage he imposes upon others or rewarded for any benefits conferred. It is generally believed that although "external (dis)economy" is also a kind of externality, it is pecuniary (monetary) externality because it will not cause the divergence between private cost and social costs, not like the direct or technological externality.

Factory pollution is often cited as a typical example of the externality that causes the divergence between private and social costs. There are costs in the factory production (such as purchasing raw materials, renting factories, machinery and equipment, paying workers' wages, etc.), and in addition, it also causes various pollutions such as air, noise pollution, and possibly water pollution, etc that damage the environment and there are also costs. However, the factory only calculates the former costs (private costs), and does not care about the costs of environmental pollutions (external costs). Therefore, the so-called "optimal" output calculated according to the equal marginal principle ($MR = MC$) is not optimal for the society (Pareto-optimal). The social costs of factory production is sum of the private costs borne by the factory itself and the external costs that it does not bear, so the social costs are greater than the private costs. This divergence between private and social costs is precisely a kind of externality: the residents nearby are involuntarily affected unfavorably by the pollutions of the factory production, and the factory does not bear the costs of the damage on the environment.

When there is divergence between private and social costs, the optimal output satisfying the equilibrium condition of $MR = MC$ for the whole society should be lower than that when only the private costs are calculated. Because the social costs (including external costs) are higher than the private costs (the MC is higher), according to the law of increasing marginal cost, the optimal output needs to decrease to make the MR equal to the higher MC . In other words, the factory's optimal output for itself is actually too high for the society, which deviates from the equilibrium condition or optimality. Welfare economics thinks it to be "inefficient" and be not Pareto-optimal. The too much output of the factory is produced at the cost of polluting the environment and harming the interests of the residents nearby, and the external costs are marginally higher than the MR of the factory. Thus, welfare economics criticizes that the free market will cause the divergence between private and social cost (externality) that is a kind of market failure, and government should intervene with it by taxing the factory so that the tax bearing are equal to the external costs of pollutions and the factory will reduce the output to the level that conforms to the social optimum (Pareto-optimality).

Externalities are not always harmful (social costs are not always higher than private costs), and may also be beneficial (social benefits may be higher than private benefits). For example, one is engaged in tree-planting and obtains income from timber. Tree-planting can make the air better, and if there is a river nearby, it is also good for conservation of soil and water. However, the planter only obtains income from timber (private benefits), and cannot get the social benefits of clear air and conservation of soil and water. Thus, there is divergence between private and social

benefits. The planter only calculates the private benefits to decide the optimal output of tree-planting according to $MR = MC$, so it will be lower than that when the social benefits are also taken into account. Welfare economics thinks it is also a kind of market failure, and government should intervene with the market by subsidizing the planter so that the subsidies are equal to the external benefits of tree-planting and the planter will increase the output to the level that conforms to the social optimum (Pareto-optimality).

However, according to the Coase theorem introduced in the previous lecture, the real cause of externality (divergence between private and social costs or benefits) is not market failure, but government failure in not delineating property rights so that the costs or benefits brought by private behaviors cannot be borne or obtained through market transactions. The problem of factory pollution has been analyzed in detail in the previous lecture. As long as the rights of air are clearly delineated, no matter to the factory or the residents nearby, the final outcome through market transactions will be the same and Pareto-optimal, which means the external costs will be internalized (also become the private costs of the factory) and there will be no divergence between private and social costs.

The case of tree-planting can be analyzed similarly: as long as the rights of air are clearly delineated, regardless of the initial assignments, the final outcome through market transactions will be the same and Pareto-optimal, which means the external benefits will be internalized (also become the private benefits of the planter) and there will be no divergence between private and social benefits.

More specifically, if the rights are delineated to the planter, those who benefit from the better environment can pay to him in exchange for more tree-planting. As long as the MR of better environment is higher than the MC of the planter, and the payment is higher than MC and lower than MR, the parties involved have no reason to refuse the transactions. Those who benefit from the better environment will be paying until the MR falls to be equal to the MC of tree-planting due to the law of diminishing marginal returns (the law of increasing marginal costs).

On the contrary, if the rights are delineated to those who benefit from the better environment, they have the rights to require the planter to continuously plant trees to improve the environmental quality, but they do not have to bear the costs of tree-planting. Now it is the planter who will pay in exchange for reducing tree-planting. As long as the MC of the planter is higher than the MR of better environment, and the payment is higher than MR and lower than MC, the parties involved have no reason to refuse the transactions. The planter will be paying until the MC falls to be equal to the MR.

In fact, Coase theorem is precisely put forward to clarify that the real cause of externality or the problem of social costs is not market failure, but that property rights are not clearly delineated. Before Coase, when externalities or the divergence between private and social costs were discussed, welfare economists always suggested government to intervene with the market, not by delineating the property rights clearly, but by taxes or subsidies. It is wrong for welfare economics in failing to understand the real cause, let alone the real solution to the problem. What is more, if there is market failure, is there also government failure? How can government know better than the market how many external costs or benefits there are, then how can it know how many taxes or subsidies it should impose or provide? If government calculates less, there is still externality. And if it calculates more, there will be new externality that may be even more serious than the old problem without government's intervention. What measurement is appropriate to determine how serious an externality is? Such ambiguity naturally opens the door for government to intervene with the market at will, which increases the power and income of government.

In other words, welfare economics and Keynes's theory have the same fate: they seem to be so reasonable to be popular, but in fact they are only made use of by government. It is precisely why Keynes's mistakes are now better known in the theoretical circle, but Keynes's theory is still adopted by many governments in practical policies. Even worse, welfare economics is still widely recognized in both theoretical circle and practical policies.

A. C. Pigou (1877-1959), who inherited Marshall's position as a lecture professor at Cambridge University, first put forward the concept of externality in welfare economics. In his "Welfare Economics" published in 1920, he gave the example of factory pollution. He also gave an example of a highway. There are two highways from city A to city B. One is short but narrow that can be called "Short Way", and the other is long but wide that can be called "Long Way". Of course, self-interested drivers will all choose to the "Short Way", so each driver will make the "Short Way" more blocked, causing the others drive more slowly. However, every driver only cares about his own cost of time, and does not care about the external cost of making the "Short Way" more blocked and causing the others' costs of time to rise, which means there is divergence between the private and social costs of time. The "Short Way" is getting more and more blocked, so at last there will be some drivers who will switch to the "Long Way". In equilibrium, it will take the same time for the "Short Way" and the "Long Way".

Pigou argued that government could force some drivers to switch to the "Long Way" by taxes without harming them. Since the time of the drivers on the "Short Way" is the same as that of those on the "Long Way", while the "Short Way" will be less blocked with some drivers switching to the "Long Way", those who continue to use the "Short Way" will be better-off. In other words, government's taxes benefit some without harming anyone, which is Pareto- improvement. Pigou also calculated the so-called ideal tax that makes some drivers switch to the "Long Way" that the private costs of those who continue to use the "Short Way" are the same as the social costs.

However, after Pigou published this example, Knight criticized it as early as 1924, pointing out that the reason of the problem was that there were no PPRs for Pigou's highways. If the "Short Way" is owned by someone, he will naturally collect an equally ideal toll, and the value and effect of it will be exactly the same as the ideal tax calculated by Pigou. In other words, Pigou thought the divergence between private and social costs implied market failure, while Knight responded that there was no market (no toll) due to the lack of PPRs, so it is not market failure, but government failure for no delineation of property rights.

To Knight's criticism, Pigou's reaction was to delete this example in the later version of "Welfare Economics". He still kept the other examples, which shows that he did not really understand Knight's criticism, or he would not admit he was wrong, so he only deleted the example criticized directly by Knight, but refused to completely abandon his "Welfare Economics" that analyzed the divergence between private and social costs with similar logic in the example of highways.

Section 2: Seemingly no Market Implies Externality?

Pigou was concerned about the divergence between private and social costs, while J. E. Meade was concerned that between private and social benefits by putting forward the famous example of bees in 1952.

Bees feed on the apple blossom without paying to the owner of apple trees. As a result, there is divergence between the private benefits of the apple-farmer and the social benefits (the external benefits are the increase in the income of the beekeeper), leading to a lower output of apple trees, which is inefficient. Government should subsidize the apple-farmer to plant more apple trees.

On the other hand, when bees feed on the apple blossom, they also fertilize the apples, but the apple-farmer does not pay to the bee-keeper. As a result, there is divergence between the private benefits of the bee-keeper and the social benefits (the external benefits are the increase in income of the apple-farmer), leading to the lower output, which is inefficient. Government should subsidize bee-keeper to raise more bees.

However, Steven N. S. Cheung conducted a field survey of bee-keeping and fruit-growing near a city with a nickname of the capital of apples in Washington State of the US in 1972. He found that bee-keepers and fruit-farmers actually sign very detailed contracts to trade pollination services and honey collections, and the pricing is as accurate as other goods in the market. Based on this survey, he published "The Fable of the Bees" in 1973, refuting the fable that has been spreading for 20 years. There are actually market transactions in the reality, but economists who just imagine what has happened outside in their ivory towers are very ignorant.

As pointed out in the analysis of factory pollution, there is transaction cost (in broad sense) for clear delineation of property rights, and if it is higher than the increase in asset value, people will choose not to clearly delineate property rights. Furthermore, even if the property rights are clearly delineated, there is also transaction cost (in narrow sense) for the market transactions, and if it is higher, neither will people choose to use the market, but the other constitutions such as the court will replace the market. Although the arbitrary judges of the court are more difficult to satisfy the equal marginal principle due to the lack of guidance of market price, resulting in lower asset value compared with the market transactions, it can be the best choice when the increase in transaction cost of market transactions is taken into account.

Similarly, there seems to be externalities in reality, because the transaction costs of delineating the property rights and market transactions are too high. With the transaction costs taken into account, there is in fact no divergence between private and social costs (benefits). In other words, externality is just an illusion caused by ignoring the transaction costs.

Wrong empirical analysis will lead to wrong policy recommendations, which has been fully reflected in the problem of externality. The analyses about externality are mostly applied to a branch called "Development Economics" that is an economics of policy recommendations. It studies which kind of investment should be encouraged by government subsidies or suppressed by taxes in order to speed up the economic growth. One of the keys is that the investments should be valued from the perspective of social costs or benefits. If there is divergence of private and social costs (benefits) in the investment, government should intervene with the market.

However, according to the above correct analysis, what government should do is to delineate the property rights, and try to effectively reduce the transaction costs (in narrow sense) in the market transactions. Obviously, it is consistent with the analysis in Lecture 22 which pointed out that the reduction in transaction costs is one of determinants of economic growth.

What is more, sometimes there seems to be no market, perhaps because the transaction cost of using the market is too high, or there may be actually be market, but the transaction cost of direct transactions is too high, so the indirect transactions take the place.

The most typical is the price of living noise and environmental quality that is often included in the rent or price of a house and indirectly traded in the market. For example, the barking of dogs and the crying of babies will disturb the residents nearby. There seems to be no market transactions for them. There are two examples in reality to show how they are transacted. One of the examples is about a primary school teacher who was always preparing lessons late at night. A neighbor living in the same building had raised a cock that crowed very early and disturbed her work and rest. She visited the neighbor and paid to buy the cock and then killed it. After the neighbor knew the purpose of her, they were embarrassed and never raised cocks again. The other example is about a family with a child. The child often dragged tables and chairs when playing, and the harsh noise from the friction with the ground disturbed the neighbor living downstairs. The neighbor visited the family with toys and said, "Does your child have no toys to play with, so he plays with tables and chairs? I have bought toys to him." Of course the child had toys, and the neighbor just politely reminded the parents to restrain the child from dragging tables and chairs. These are direct but "clever" transactions, and the payments for the cock and toys are actually the payments to buy the right to be free from noise.

There are also indirect transactions. For example, high-class apartments abroad often have the stipulation that families with pets or children are not accepted. The rent for these apartments is higher than those without such a stipulation, which implies the rent includes the price of "buying" the right to enjoy a quiet environment. More commonly, if the green is good in a community, or the community is located in a beautiful natural environment, the price of the houses must be higher, which implies the higher price includes the payments of the right to enjoy high-quality environment.

The economists who do not notice all these indirect transactions and think there is market failure again are no better than Mead who even did not know there were actually market transactions of pollination services and honey collections.

Furthermore, sometimes there is no market transaction, probably because there is no divergence between private and social costs at all instead of due to market failure! Steven N. S. Cheung has given such an example to illustrate this possibility: one plays the piano at home, and the sound can be heard by his neighbor. At the beginning, the neighbor enjoys the sound of the piano, but he does not pay to the pianist, so there is divergence between private and social benefits. As the pianist keeps playing, according to the law of diminishing marginal returns, the MR of the sound of the piano will fall and even becomes negative after a certain point. In other words, the sound of the piano has changed from pleasant music to disturbing noise for the neighbor, but the pianist does not pay (compensate) to him, so there is divergence between private and social costs. However, if the pianist stops playing precisely at the point where the MR falls to be 0, the neighbor needs not to pay him and neither needs he compensate the neighbor, so there is no divergence between private and social costs or benefits.

Some may wonder: how can the pianist be so precise. Firstly, the focus of this example is that the lack of market transaction does not necessarily mean there must be divergence between private and social costs, but there may be no need for market transaction. Therefore, it is logically wrong for economists or government to regard the case without market transaction as that with externality. Secondly, if the pianist cannot be so precise, why can government be so precise to

know how much tax or subsidy should be imposed or provided to satisfy the principle of $MR = MC$? Why can government as a third party outside of the market know better than the parties in the market?

Section 3: The Externality from the Perspective of Contract

From the perspective of contract, the essence of externality will be clearer. Externality implies that one's behavior affects the others, but there is no contract to restrain (the bad effects) or buy (the good effects). The causes of externality can be classified into three categories: (1) there is no delineation of property rights, so there is no contract for market transactions; (2) there is a contract, but the use clauses are incomplete; (3) there are complete contract clauses, but MR and MC are not equal for some uses.

As mentioned in Lecture 14, there is not only price clause (income clause) but also use clause in a complicated contract, which means the contract is structural. And as mentioned in Lecture 20, the properties of direct measurement are priced by the price clause in a contract, while those without direct measurement are supervised by the use clause. Lease contracts (including the employment contracts for labor, because in a society where slavery is not allowed, labor can only be rented but not purchased) are all structural. By contrast, in buy-out contracts there is usually only price clause, so they are not structural and much simpler.

As mentioned in Lecture 23, there are three rights in PPRs. Lease contracts only deal with part of the rights of assets (usually use right and right to income for a period of time). The lessee's use often affects the value of the leaser's asset, which leads to divergence between private and social costs. For example, careless use will shorten the life of an asset, but as long as the life does not become even shorter than the lease period, it will not affect the lessee but will affect the leaser. Thus, lease contracts need to restrict the behavior of the lessee from careless use with the use clause to prevent divergence between private and social costs as much as possible.

Case (1) is a kind of government failure. Even if there are contracts of market transactions it is still possible for divergence between private and social costs such as Case (2) and (3). For Case (2), there is no contract in the world that can cover all the possibilities about the use of assets in detail, because the transaction cost is too high. On the other hand, besides the clauses written explicitly in a contract, there are laws, default conventions, customs, morals, etc that are not necessary to be repeated in the contract can restrict the lessee. For Case (3), under the constraint of transaction cost, the increase in asset value brought about by equalizing MR and MC is too negligible to outweigh the transaction cost.

In a word, there are following mistakes in the analysis on externality in MSE. (1) There is no market transaction, probably because there is no divergence between private and social costs at all and no need for market transaction. That there is no market does not mean externality. (2) There is no market because there is no delineation of property rights. According to the first version of Coase Theorem, without the prelude of delineation of rights, there cannot be market transactions. (3) There is delineation of property rights, and there are also market transactions, but because the transaction cost of direct transactions is too high, indirect transactions take the place. Sometimes, because there have been implicit contract clauses such as laws and customs, the relevant restrictions are not repeated in the explicit contract clauses. (4) There is delineation of property rights, but the transaction cost of using market (transaction cost in a narrow sense) is too high, so the non-price criteria (non-market institutions) such as judges' arbitrary decisions (the court) are used to deal with the problem of externality.

Lecture 25: Rent Dissipation

In this book, the concept of “Rent Dissipation” has been used to analyze government’s intervention with the market in Lecture 9, pointing out that the essence of price control is the infringement on right to income right of the seller. Price control causes the part of the income (rent) to become ownerless, and self-interested people rush to compete for it, resulting in rent dissipation.

Price control is criticized by welfare economics as inefficient or wasteful. Having learned Coase theorem, readers should understand that the reason lies in that the right to that part of income is not clearly delineated, so there is no market transaction for it, which means it is impossible to use price criterion to restrict competition or determine the outcome of competition, so non-price criteria take the place that must lead to rent dissipation.

Although price control has been analyzed in detail in Lecture 9, the concept of “rent dissipation” has not been systematically introduced, which will be the theme of this lecture.

Section 1: The Concept of Rent Dissipation

The example of Pigou's highways mentioned in the previous lecture can be analyzed from the perspective of rent dissipation. If the "Short Way" is delineated as PPRs, the owner can obtain income (rent) through market transactions, but in Pigou's example it is public, so the competition of drivers leads to congestion, increasing the driving time and causing the rent of it to fall to zero. In other words, the MC of driving time is equal to the tolls if there is an owner, but the rent only dissipates when the "Short Way" is public. The costs of drivers are actually the same, just when is an owner, they pay the costs in the form of tolls that become the income (rental) of the owner, while when there is no owner, the costs of driving time increase to offset the rent that does not become anyone's income but dissipates.

Later, an economist named H. S. Gordon extended the example of highways to that of high seas, and put forward "the Tragedy of the Commons". The rent of common goods that are not delineated as PPRs will drop to zero due to too many competitors and overuses. The term "rent dissipation" was first put forward by him, but there are mistakes in his analysis. Steven N. S. Cheung has given the correct analysis with an example of a lake in his "Economic Explanation".

The analysis can be presented by the Fig.20-2, but the general production factor specifically refers to the labor input of fishermen, so the horizontal axis of Fig.25-1 is L, and the vertical axis is the price of labor (W) that is the highest wage the fishermen have given up in other industries.

If there is an owner of the lake, he will restrict the total fishing time (the number of fishermen multiplied by the fishing time per capita) to L^* where the equilibrium condition of $mrp = W$ is satisfied, and the total fishing income is WEL^*O , while the rent of the great lakes is the triangle above WE.

If the lake is ownerless or a common good, without the restriction of the owner, people will flock to fish, and one more fisherman join in will cause the others' mrp to move a little toward the lower-left because he will occupy the others' fishing spaces and opportunities. Thus, the equilibrium point (E) of $mrp = W$ will move along the horizontal line of WE to the left, which means L^* is decreasing or the others are reducing a little fishing time¹. As a result, the triangle area above WE is constantly decreasing, which means the rent of the lake is constantly decreasing. More and more people join in fishing, and everyone's curve of mrp is moving towards the lower-left until the equilibrium point (E) approaches the point W infinitely (or the curve of mrp moves to mrp') where everyone's fishing time approaches 0 infinitely and the triangle area above WE drops to nearly 0, which implies all the rent of the lake dissipates.

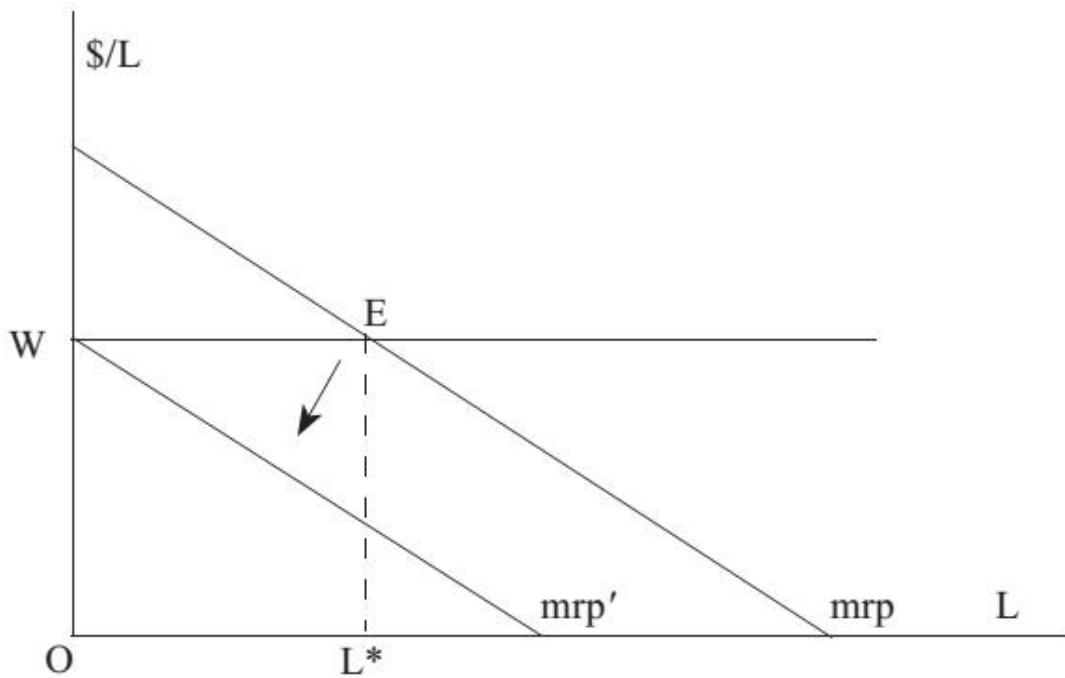


Figure 25-1

In reality, it has never been seen that countless people compete for the use of common goods, neither their participating time close to 0, because different people have different costs of time, skills and information costs that lead to different mrp and W . When some's curves of mrp have moved to mrp' , another's are still at the upper-right, so the latter's participation time will not be as low as 0, but the triangle above the WE for them does not represent the rent of the common good (such as the lake), but the rent obtained by them due to their lower costs of time, or better skills, or the better information. On the other hand, in reality, government often restricts competition by limiting the number of participants (such as by license), so that the rent of common goods will not dissipate to 0.

Thus, the concept of rent dissipation can be deduced based on the above analysis. When the rights of an asset are not delineated as private, people will compete to use it with non-price criteria, resulting in a decline in the income (rent) of it.

Attention: although the rent dissipation of highways or high seas (the lake) is caused by too many users in the Pigou's, the Gordon's and the Cheung's examples, in reality, there are other cases of rent dissipation that the assets are used to produce lower rent, which also leads to the decline in rent. For example, there is a pasture in Africa, where the climate and soil are suitable for planting almond trees with high value. However, because it is a common good, no one will plant almond trees there, but use it for grazing cattle and sheep. Cattle and sheep are private property, and can be driven home at night, while almond trees cannot be prevented from cutting off and stealing away.

1. The number of fishermen cannot be reduced because there is no owner to restrict [the use](#)

Section 2: The Relation between Non-Price Criteria and Rent Dissipation

According to the concept of rent dissipation, there is no rent dissipation for price criterion, while there is for non-price criteria. Why? If price criterion is adopted to restrict competition (determine the outcome of competition), the owner of the asset can obtain the rent (the price paid by the users), so the rent will not dissipate, and the legal system of protecting PPRs behind the price criterion also ensures that the price paid by the users is the income they obtained through providing products or services, which means they must make contributions to the society first.

By contrast, if non-price criteria are adopted, there will be more or less rent dissipation. For example, if it is the criterion of first-come-first-served, people will have to wait in a queue and the time spent does not contribute to the society, nor the suppliers of the goods obtain any income. The cost of time in waiting is equal to the price that people would have paid if price criterion is used, so the rent dissipates.

If it is the criterion of force, people will invest in weapons, which will inevitably cause the scarce resources to be diverted from production to weapons, resulting in a decline in the output of the society. On the other hand, bloodshed and destruction caused by fighting also reduce social wealth. It is said that during the “gold rush” in California, the criteria of first-come-first-served and force were mixed to determine the ownership of the discovered gold, so people raised the sled dog very strong in order to reach the scene as quickly as possible. If there were still many people arriving at the same time, they would fight with each other.

If it is the criterion of seniority, which means the older can get higher income, one will choose to just sit there doing nothing to wait for becoming older, so his contribution to the society will be marginally lower than the cost paid by the society to him. He would have increased his contribution if he worked harder instead of waiting to become older, so the rent of his labor has dissipated.

If it is the criterion of hierarchy, which means the superior has the right to decide the distribution of wealth, people will be keen on developing interpersonal relationships such as flattery, which is not helpful for increasing production and also result in a decline in the output of the society.

What is the relationship between Coase theorem and rent dissipation? According to the first version of Coase theorem, the delineation of right is an essential prelude to market transactions, while the essence of market transaction is precisely price criterion. There will be rent dissipation when there is no owner of assets, so the outcome of competition does not determined by price criterion (price is income or rent from the owner's perspective), but non-price criteria that lead to the decline in the value of the assets. Thus, if non-price criteria are adopted in reality, it is probably due to the lack of delineation of rights.

Of course, there is transaction cost in reality. Even if there is delineation of rights, the transaction cost (including information cost) of price adjustment may cause seemingly inequality of supply and demand temporally, so there will be the phenomena of non-price criteria such as “queuing”. As mentioned in Lecture 7, there are different fluctuation frequencies of price in the supermarket and the farmers market, which is an example. It is the owners of the vegetables who directly sell them in the farmers market, while it is the front-line employees who are not the owners in the supermarkets, so the delineation of the rights is somewhat vague in the latter case. The information about the change in supply and demand needs to be passed to the owner (the boss) of the supermarket, and then the instruction of price adjustment also needs to be passed to the

front-line employees, so there is obviously higher transaction cost (information cost and the cost of price adjustment) in the supermarket than in the farmers market.

On the other hand, as mentioned in Lecture 17, the serious queuing when a new generation of iPhone is for sale is due to the lower price than the equilibrium voluntarily and deliberately charged by Apple Inc. Although there is rent dissipation for this non-price criterion, but the queuing has an effect similar to advertisement, which is beneficial to Apple Inc. In other words, even if there is clear delineation of rights, people may still choose to adopt non-price criterion voluntarily, because the rent dissipation is possible to exchange for the reduction of transaction cost. According to the equal marginal principle, the rent dissipation caused by non-price criteria should be marginally equal to the transaction cost reduced.

Section 3: The Core of the Theory of Price Control

Price control is one of the main causes of rent dissipation in reality, which has been analyzed in detail in Lecture 9.

The analysis of price control in MSE focuses on the so-called "shortage", but the key lies in the fact that price control partially infringes on the right to income right of the seller, which leads to non-price criteria. Take rent control as an example. Suppose the rent in the market is RMB1000 per month, and is controlled to only RMB800, so there is RMB200 of income that becomes ownerless income. Government only forbids the owner to collect this part of rent, but does not clearly say who has right to it. If it is owned by the lessee, the use right is actually in the hand of the owner, and he can take it back by many means such as contract fee. Even if government strengthens regulation and successfully prohibits him from taking back the regulated rent, the owner can degrade the quality by not providing furniture and appliances, and even stops supplying. Anyway, the owner has a lot of means to prevent the lessee from getting this part of income.

Nevertheless, if government clearly stipulates that the value of the property is divided by equity, 80% belongs to the original owner and 20% belongs to the lessee, the latter actually becomes the owner of the 20% of shares. In this case, the property rights are clearly delineated, and the so-called lessee is similar to the manager with shares in a company. On the one hand, he actually runs the company (uses the house), on the other hand, he also has part of the shares. He distributes 80% of the income (RMB1000) to the original owner (RMB800), and keeps the other 20% (RMB200). Obviously, a manager has some shares of the company, which will not cause any rent dissipation, but will care more about the company because some of his interests are closely related to the operation of the company. Now, don't many companies offer equity to the executives as motivations?

However, in reality, government that implements price control never does so. The regulated price becomes ownerless income, which causes self-interested people to compete by non-price criteria and so rent dissipation. Why does not government do so to avoid the loss of rent dissipation? Because if government does so, it will become very obvious that price control is equivalent to openly robbing the owner of the property rights and transferring to the lessee. Although the property rights are clearly delineated after the robbery that will not cause rent dissipation, the robbery itself is the infringement on property rights and will make the owner take defensive measures that will cause rent dissipation. Government frequently does such things as openly robbery, which is equivalent to poor public security that thieves can openly rob without being stopped by police, resulting in the transfer of property rights from the original owners to the thieves. In such a society, people will not be willing to have property or engage in production, and the economic performance will naturally be poor. Property rights are not delineated in fact because they cannot be protected. That is why in reality, price control never starts with the delineation of property rights, but the infringement on the right to income, and government tries to help to ambiguously delineate the rights by laws ¹.

As mentioned in Lecture 14, the purpose of all institutions is to save transaction cost, and the market is also an institution, so it is also used to save transaction cost. The question is: what kind of transaction cost is saved by the market? According to the analysis of rent dissipation, what the market saves is mainly the transaction cost of rent dissipation. Because using the market means using price criterion to determine the outcome of competition, and price criterion is the only competition criterion that will not cause rent dissipation.

Of course, price in the market can also transmit the information such as the changes in supply and demand or the quality of goods to various economic entities in the market with low cost, which means the information cost is relatively low. However, the behaviors of price-making analyzed in Lecture 17 shows that price in the market can also be used to mislead consumers, which lead to the increase in information cost. Thus, the market as an institution can save some transaction costs, but will also lead to the increase in other transaction costs. People will compare the costs and benefits of this institution to decide whether to use it or not.

The core of the theory of price control is not that there will be competition by non-price criteria for ownerless income, which will cause rent dissipation, but that there are so many non-price criteria, and which one or combination will be chosen? Due to the postulate of self-interest, rent dissipation is bad for anyone, so all will try to reduce rent dissipation as much as possible under certain constraints. In other words, the rent dissipation under price control should be minimized under the constraints. There are two kinds of effects for price control: one is the rising transaction cost, and the other is the falling asset value. Therefore, people will choose the non-price criterion that causes the least increase in transaction cost and the least decrease in asset value.

Take the example of price control of petroleum in the 1970s in the US mentioned in Lecture 9 again. Why did not gas stations take back the regulated price directly by contract fee, just like the owners of house who were imposed by rent control, but indirectly did so by increasing the price of lubricants that were not subject to price control? The reason is obvious: refueling is not a one-time sale within a long time (such as one year) like renting house, so the transaction cost caused by negotiation about contract fee is higher. If the gas stations, like the owners of house, negotiated with everyone who came to refuel (those who accepted contract fee were given the priority to refuel, and those who did not had to wait in line), there would still be a long queue because the negotiations inevitably took some time. In other words, due to the different constraints of refueling and renting house, the transaction cost of contract fee for taking back the income infringed by price control is different, and the rent dissipation is also different. Self-interested people will choose the means with least transaction cost and least rent dissipation.

However, when government thinks of “strengthening regulation”, it will first press such countermeasures with least increase in transaction cost and least decrease in rent dissipation, which naturally causes greater social loss.

1. If the delineation is too clear, it will be too obvious that government is openly robbing the original owners, so there will still be ownerless income, which attracts people to compete by non-price criteria, resulting in rent dissipation. [↗](#)

Section 4: Rent Dissipation as a Mode of Thinking

Rent dissipation is not only a phenomenon, but also can be used as a mode of thinking. The ownerless income will dissipate because self-interested people will compete to own it. There is a Cantonese saying that "There is impossible to be such a big frog on the street", which implies that if there is a big frog on the street, people will naturally try to catch it to cook. It is similar to the saying "Money doesn't grow on trees."

As a mode of thinking, there are two uses for rent dissipation.

(1) If there seems to be ownerless income, there must be some hidden constraints behind that imply there is actually no income or the cost of obtaining the income is higher than it. There is a true story in ancient China that can best illustrate this point. Once some children found that there were full of plums on an ownerless tree by the roadside. They all ran to climb the tree to pick the plums except one who said, "These plums must be sour." After the other children picked the plums and tasted, they found indeed so. They asked the smart child why he could know in advance. He answered, "If the plums were sweet, would not they be picked long ago?"

From this ancient story to the reality nowadays, mango trees are planted as green trees in many southern cities, which attract many people to climb the trees to pick them when the fruit season comes. In fact, these green trees only bear green mangos that are small and sour compared with the mangos as fruit. Even so, as long as there is such a little profit, people will take the risk of falling from trees to climb and pick, so that some cities consider replanting other green trees without fruits.

This use of rent dissipation helps the economists who are engaged in economic explanation or empirical test to realize that there may be constraints they do not notice.

(2) If there is ownerless income in an economic analysis, it must be wrong. It is a quick and accurate way to judge whether a theory is right or wrong. A theory that is mathematically proved is only right in logical reasoning from the premise to the conclusion. However, because there is not content in mathematics, it is impossible to guarantee or verify the economic implication of the premise and conclusion. By contrast, as a mode of thinking, there are rich economic implications for rent dissipation that are directly related to the postulate of self-interest. Therefore, an analysis that is wrong seen from rent dissipation must be wrong in the economic implications against the postulate of self-interest.

Attention: That the ownerless income will be dissipated by competition means that it will drop to zero marginally, not that it will all dissipate, because there will still be income (rent) within the margin that cannot be dissipated by competition. Take the example of fishing on the high seas, anyone can fish, but if one has particularly better fishing skill and will get much higher fishing income that cannot be dissipated by the other competitors, so he is the owner of the higher fishing income (rent). Strictly speaking, the rent is the return for his superb fishing skill that is unique to him, and others cannot overcome this threshold to compete with him without learning or talent, so the higher fishing income is a monopoly rent.

Lecture 26 Political Economics

In the period of classical economics, in order to avoid mistaking economics for family financing by outsiders, classical economists added "political" before "economics" to emphasize that it was state financing that economics studied, which is in fact not the scientific nature of economics. When Marshall wrote his "The Principle of Economics", the term of economics has become well-known, so nobody would mistake it. He thus canceled the adjective of "political". There are also many outsiders nowadays who mistake "political economics" in classical economics for the economics taking the political factors into account, but it is the same as economics.

However, the mainstream economics does have been just focusing on the economic activities in the market, or the choices based on price criterion. Some economists apply the economic analysis to other fields of social science after the scientific nature of economics became clear. As a result, there is so called "Economic Imperialism". Especially after the concept of transaction costs were advanced, it has become common to apply economics to analyze the political phenomena. So the term of "Political Economics" is reborn as a cross-discipline of economics and politics, but it is different from the meaning of this term in classical economics.

Section 1: The Nature of Nation from the Perspective of Economics

The traditional economics focuses on individual as the starting point to study human's behaviors: how a self-interested individual chooses the most beneficial behavior under specific constraints. Although the market is formed by the joint behaviors of many individuals, and although the so-called macroeconomics needs to sum up all individual choices in an economy, they are all based on individual's choice. Economists notice that the area where market (price criterion) does not apply is the public domain. The choice in the public domain is called public choice.

For example, it is the owner of a house to decide how to allocate the spaces in his house, such as how big the bedroom, study and hall are respectively, and how to place the beds, tables and chairs. If he has exclusive ownership of the house, then it is his Robinson one-man world and there is no market within it because there is no society. However, if he chooses to rent the house to more than one tenant, each of whom rents a part of the house, there is a market within it. If a tenant wants to increase his living space, he can negotiate with the owner (landlord) of the house and pay more rents in exchange for the increase in space.

When multiple tenants rent a house, it is inevitable for them to share some spaces, such as kitchen, toilet, bathroom, passageway, etc. All tenants need to jointly decide how to allocate the use of these shared or public spaces, and undertake the cleaning of them. The landlord may take the responsibility to formulate a convention which is written in the lease contract, requiring all tenants to obey. He may be also responsible to clean the public spaces, but the rent he charges will definitely be higher. Here the landlord has actually become a government. The convention he formulates is equivalent to laws, and tenants who do not obey it will be deported by him. The higher rent he collects is equivalent to tax. As mentioned in the previous Lecture 23 about Coase Theorem, "rent" and "tax" in ancient times were often synonymous.

Of course, the tenants may negotiate to formulate a convention themselves, and each is assigned the responsibility of cleaning the public spaces in turn. There are not too many tenants in a house, and neither are there too many public spaces. Since everyone lives under the same roof, it is convenient to jointly make decisions by all together sitting down and having a talk to reach an acceptable agreement. Here it is direct democracy where all participate in making decisions on public affairs.

Let's extend the analysis of multiple tenants in a house to multiple owners in a building or community. There are much more owners now, and they live in their own houses, so it is far less convenient to jointly make decisions than the tenants living under the same roof. Moreover, there are also much more public affairs than before, such as the security measures for the building or community, the maintenance of elevators, lights, greenings, and the color of the building's exterior walls to be painted and maintained ... Under such circumstances, owners committee is often set up to handle these public affairs. What's more, because there are too many owners with different opinions, it is difficult even impossible to reach an agreement that everyone accepts, and a vote needs to be taken to decide. Obviously, here it is indirect democracy where the owners committee is equivalent to the government or parliament.

In reality, buildings or communities are often managed by property management companies (PMCs) which have more professional knowledge and skills than owners. Here PMC is equivalent to a government, and the property management fees it collects are equivalent to taxes. If there is no owners committee in a building or community but only a PMC which designated by the developer instead of the owners, the government is autocratic. However, it does not mean that

the owners cannot get rid of a PMC that they are extremely dissatisfied with. It is not unique in reality that the owners unite to drive out the original PMC and hire a new one, which is equivalent to an uprising to overthrow the existing government that people are extremely dissatisfied with, and establishment of a new one. If there are both an owners committee and a PMC in a building or community, the former is equivalent to parliament while the latter government.

It can be concluded from the above examples that a nation is only an expanded community. Every owner (people) has his own house (private property). Whoever wants to increase his living spaces can buy from another owner through market transactions. In other words, the private affairs such as the use of private spaces are handled by individuals through the market. However, it is PMC (government) that deal with the affairs in public spaces which should not intervene with the affairs in private spaces. Owners committee (parliament) should represent the owners to make public choices on the affairs in public space. Unless there is a proprietor (dictator) who has clear PPR of the whole community just like the landlord of a house, the PPR of the public spaces actually belongs to the owners. In theory, PMC should manage the public spaces in the community according to the will of owners' committee.

In brief, there are three ways to make decisions (choices) on different affairs: market, dictatorship and democracy. Among them, the market is used to deal with private affairs where the PPRs are well delineated, while Dictatorship and democracy are used to deal with public affairs where the PPRs are not well delineated, so the properties are shared or publicly owned.

Section 2: The Definition of Dictatorship and Democracy

The Market has been analyzed a lot in the previous lectures, so this section focuses on these questions: what is dictatorship? And what is democracy?

Ruling out subjective feelings and ideologies that are incompatible with the requirement of objectiveness in science, “dictatorship” literally means deciding alone. Deciding alone does not mean refusing to listen to other’s opinions. A dictator may or may not listen which is his own choice. What matters is that he is the final decider. In theory, he does not need to obey or submit to the will of others. In fact, on private affairs, each of us is a dictator because how to use those we have clear PPRs is, of course, entirely up to us to decide alone.

Extending the logic above, the boss of a company must be a dictator. He is equivalent to the landlord of a house mentioned in the previous section who decides alone how to use the public spaces in his house. What about the manager in a company? He is hired by the boss (owner) of the company. Within the authority stipulated in the contract, he is also a dictator. Although the boss is the real owner of the company and has the right to fire the manager when he is dissatisfied with his decision, as long as he is still in office, according to the authority given by the contract, the manager does not need to submit to the will of the boss in theory. Similarly, PMC of a community is also a dictator within the authority stipulated in the contract. Its management of the residential area must be obeyed by every owner in the residential area. If some owners are dissatisfied, they can revise the authority through owners committee, or even change the PMC (just as the boss of the company can revise the contract or even fire the manager if he is dissatisfied with his performance). However, if they fail to do so through owners committee, they still have to accept the PMC’s decisions on the public affairs. In the previous section, it has been pointed out that PMC in a community is equivalent to a government, so government is also a dictator during its term of office.

In modern politics, there is system of separation of powers, among which government, which is a dictator, has the executive power, and court, which is also a dictator, has judicial power. Unless using the jury system, judges decide alone on the cases. A government is likely to form by election, while judges are not even elected, so their autocratic style is even stronger. Some will certainly argue that judges decide according to laws and facts, so they are not dictatorial. But which dictator in the world does not decide alone based on something? A completely unfounded decision is not a choice but a random draw.

As mentioned above, any individual’s decisions on those they have PPRs are dictatorial, but in Lecture 23 about Coase Theorem, it has been pointed out that one has PPRs does not mean the use right is completely unconstrained. For example, if one has an apple, he can choose to eat it, give it to others or sell it, but he cannot throw it on other’s head. Is there any essential difference between that a judge must decide according to laws and facts, and that one must use an apple of his according to the laws? A judge who does not judge a case according to laws and fact will be rejected by the higher court, one who throws an apple on other’s head will be punished by law, a PMC who makes the owners dissatisfied will be expelled by the owners, a manager whose performance is poor will be fired by the boss, and a dictator who runs the nation by tyranny and arouses resentment and anger will be overthrown by people’s uprising ... Are there essential differences? The answer is no! The relevant constraints are different, and the constraint forces are different, but the nature of deciding alone is exactly the same.

Let's turn to democracy. What is the definition of democracy? There are many definitions in politics most of which are vague and cannot be tested by facts from the perspective of science. It can only be defined as voting if it is clear enough for test. Certainly many will doubt if democracy can be simply equivalent to voting, but let's check the other common definitions of it.

If democracy is defined as popular sovereignty which means the political power of a nation belongs to people, "people" itself is a very vague concept. And what is sovereignty? If it refers to the well delineated or effectively protected property rights, as mentioned in the previous examples of a house and a community, public choice is related the affairs in public spaces which cannot be well delineated as PPRs. In other words, democracy in politics is logically similar to public ownership in economics in that everyone has ownership means no one has it. It is said that a nation's properties are owned by all people in public ownership, which are actually owned by no one. Similarly, it is said that the political power of a nation is owned by all people in democracy, which are actually owned by no one. The latter analysis of democracy changing from "tyranny of the majority" to "tyranny of the minority" is deducted from the fact that democracy is essentially public ownership in politics which causes vague property rights of political power.

Or if democracy is defined as majority rule, how to identify the majority? Of course by voting. Therefore, the clearest and most testable definition of democracy is nothing but voting.

After the definition of democracy has been clarified, the definition of dictatorship also become very clear, although the previous definition of dictatorship as deciding alone is clear enough. Since democracy is defined as voting, dictatorship can be defined as not voting. From the similar perspective, the market can be defined as voting with money because it is price criterion that is used in the market.

Section 3: Tyranny of the Majority

The concept of Pareto-optimality has been introduced in Lecture 23. Regardless of the transaction cost of the institution, which can be Pareto-optimal, the market, democracy or dictatorship?

Market transactions are voluntary and mutually beneficial. Even if there is cheating, under the specific constraints (including ignorance caused by information cost), the interests of each individual are protected by his self-interested nature and the legal system of protecting (delineating) PPRs that supports the establishment of price criterion. In other words, the interests of all individuals can be expressed and protected in the market system, so it is Pareto-optimal.

However, the choices in the market are private choices. For public choice, how can the interests of all individuals be expressed and protected?

Review the previous example. Multiple tenants decide the use and cleanliness of public space through negotiation. As long as they can reach an agreement, everyone's interests must be expressed and protected by the self-interested nature, which is direct democracy. Although the specific means is not necessarily voting, if voting is adopted, since an agreement is reached, the result must be a unanimous vote. But what if an agreement cannot be reached? It means that if there is a vote, at least one casts opposing vote. In other words, a unanimous vote under direct democracy will be certainly Pareto-optimal that means no one's interests are harmed.

However, in politics of reality, it is difficult to achieve the result of unanimous vote, or else market transactions where all parties involved in the market have to reach an agreement can have been used. Precisely because the transaction cost of reaching an agreement by all is too high, people give up market transactions and turn to the principle of the minority obeying the majority. If public choice is made according to the voting result, the interests of the minority who cast opposing votes cannot be expressed, nor can be protected, which is precisely the reason why democracy will cause tyranny of the majority. In fact, this problem is not as simple as what is mentioned here, and more will be discussed later.

If it is just the case of renting a house, the minority can choose to move away, or rent the whole house by paying more rent to eliminate public space. In other words, although there is problem in the use of public space of renting in groups, which involves public choice without market, there is still a market outside the house. Voting (democracy) will inevitably harm the interests of the minority unless it is unanimously, but as long as there is a market outside, the minority can always vote with their feet or with money to protect their interests.

Similarly, if the use and cleanliness of public space is decided by the owner of the house, on the one hand, the owner's interests are directly related to the tenants'. If the use of public space is not appropriate, and the tenants feel that their interests are harmed, they will require a rent reduction or simply move away, thus reducing the owner's rent or income. On the other hand, there is also a market outside the house, and tenants who feel that their interests are harmed can choose to move away. Therefore, there may be wrong decisions through the arbitrary dictatorship of the owner instead of the market and misjudgments of the tenants' interests (and ultimately the owner's own interests) due to the lack of guidance of the price, but as long as there is a market outside, the tenants who are harmed can always vote with their feet or with money to protect their interests.

Turn to the case where the owners of a community authorize a PMC to manage the public space. Owners committee is equivalent to a parliament, and the members in it are not all owners, but elected from all owners by voting, and then they vote to make decision. Unless it is a matter that greatly affects the interests of the owners, the general daily affairs are decided by PMC. A PMC signs a service agreement (contract) with the owners committee that grants and define the authority of it, so the service agreement is equivalent to the constitution of a country. Then, the PMC also formulates a series of conventions on the use of public space in the community approved by the owners committee, and all owners are required to abide, so the conventions are equivalent to the laws of a country.

Thus, the PMC will manage the community dictatorially according to the conventions within the authority granted by the service agreement. Even if there will inevitably be some cases unfavorable to individual owners and seemingly harming the interests of them, since the PMC is authorized by the owners committee and the relevant conventions have been approved, the harm can be regarded as acceptable. If they think the harm is actually not acceptable, they can express their appeals through the owners committee, which is equivalent to the boss of a company asking the manager to make corrections. Thus, the focus of the problem is whether the owners committee can fully represent the interests of every owner.

For example, if there is a convention of prohibiting dogs from excreting at will, the owners' right to use the public space to walk dogs is obviously greatly restricted. And if most owners in the community have dogs, since the members in the owners committee are elected by voting, they may think the benefit of restricting the excretion of dogs in public space is higher than the cost of cleaning. Therefore, with the majority rule, they will approve to abolish this convention. A few owners who do not raise dogs have to bear the cost of cleaning dog excrement, but they do not enjoy the benefit of walking dogs, so their interests are harmed. Obviously, there is a divergence between private and social costs, because the property rights of the public space in the community are not clearly delineated, and the owners decide the use of the public space not through market transactions but voting. In other words, the members of the owners committee are elected according to the majority rule, so the interests of every owner cannot be fully represented, which is precisely the problem of indirect democracy.

However, there is a market outside the community. If a minority feels that his interests are seriously harmed and he is unable to change the situation under the one-man-one-vote election, he can move away from this community to a community where few have dogs, so there is the convention prohibiting dogs from excreting in public space at will.

If the votes are allowed to be for sale, and the interests of the minority who do not have dogs are more than those of the majority, the minority will pay to buy the votes from the majority. In this case, the votes have actually become transferable shares, or the property rights of public space have been clearly delineated to all owners in the form of shares (one share for one vote), so there can be market transactions, and the institution of voting has become the voting with money. In theory, with market transactions, everyone's interests can be protected. However, the problem is that the principle of determining the outcome of vote is still the majority rule, so those who buy the votes will not buy all the votes, but only buy enough votes to ensure they can win. And they will not choose to buy those who strongly oppose them, because they need to pay a higher cost to buy these opponents' votes, but the interests of these opponents are harmed the most seriously! It is precisely the reason why votes are not allowed to be for sale in democracy.¹

What is more, the representatives in the owners committees are also self-interested, so they may vote not according to the interests of all owners, but to their own personal interests. When public affairs are complicated, it is not easy to understand what the interests of most owners are, so the representatives in the owners committee can cheat the other owners with sweet words (just as sellers in the market make use of their information to cheat buyers) and serve their own interests

under the banner of serving the interests of all. Therefore, democratic voting, which is originally intended to satisfy at least the interests of the majority (although at the expense of the interests of the minority), will probably degenerate into serving the interests of the minority under the banner of representing the majority.

No wonder the political philosophers (such as Rousseau) opposed indirect democracy. Although the direct democracy they advocated is impractical in modern society with a high degree of specialized division. The political philosophers who advocated indirect democracy pointed out that direct democracy would lead to tyranny of the majority, but in fact the key is not whether it is directly or indirectly democracy, but whether the division between private space and public space is correct and clear, which will be discussed later.

Extended to the politics in reality, it is not easy to change a government, and it is also difficult to vote with feet by leaving a country, and there is no market outside a government or a country. Thus, the harm of democracy to the interests of the minority is an inevitable cost.

1. Votes are allowed to be for sale, which is equivalent to returning to the price criterion where there is no rent dissipation, so it has an overwhelming advantage over the non-price criterion that votes are not allowed to be for sale, and attracts self-interested people to try to detour and return to the price criterion, just as what people will do under price control. Therefore, the democratic countries actually allow votes to be for sale in disguise such as political contribution and log-rolling. [↩](#)

Section 4: Arrow's Impossibility Theorem and the Median-Voter Theorem

As mentioned in the previous section, the market can ensure that everyone's interests are expressed and protected, which is Pareto-optimal, but democracy cannot and can only seek to ensure that the interests of the majority are expressed and protected at the expense of the minority. Attention: the interests of the minority may not be lower than those of the majority. If the minority can buy the votes of the majority, in this case they will pay to the majority and still have surplus, which is precisely the implication of the third version of Coase theorem.

However, does democracy really at least ensure that the interests of the majority are expressed and protected? Arrow's Impossibility Theorem or the Cycling Paradox has broken the myth of democracy.

Arrow's impossibility theorem points out that a fundamental defect of the majority rule is that there will be cycling in decision-making. Suppose the voting preferences of three voters A, B, and C over three proposals, ranked from highest to lowest, are as follows: A prefers X is over Y, and Y over Z; B prefers Y over Z, and Z over X; C prefers Z over X, and X over Y.

Notice that a majority (A and C) prefer X over Y, a majority (A and B) prefer Y over Z, but a majority (B and C) prefer Z over X! This is the paradox. The majority-rule choices can be intransitive.

Where cycling is possible, the voting sequence (the "agenda") can determine the final outcome. Suppose A can determine the order of voting. He could call first for a vote between Y and Z, the winner to be matched against X. Policy Y would win the first vote. But in the second vote, his favored policy X would win over Y. Given the same power over the agenda, B or C could also assure outcomes in their own favor.

Is it possible to design a voting without cycling? Arrow's conclusion is that there is no voting that can guarantee efficiency, protect everyone's interests without relying on the majority rule. In other words, when the ranking of the interests of all in a society is known, it is impossible to rank the social interests by ranking the individual's interests, and it is impossible to accurately express everyone's interests by a certain procedure, so it is impossible to make a desirable public choice.

The above is the case where there are multiple alternatives and there is cycling. What if an issue is to be decided in dichotomous (yes/no) voting? Here it is the Median-voter Theorem that applies. Under reasonable assumptions the outcome will coincide with the preferences of the median voter. To see why, imagine lining up the voters in order of their support for the issue. The voter in the middle (the median voter) must necessarily be among the majority. If there is a strong preponderance on one side, by definition the median voter will be in that group. Or, at the extreme, if the other voters are equally split, then the middle voter's own choice makes the majority.

This median-voter theorem seems to make little sense in politics of reality because it is generally impossible for the other voters to be equally split, but it is actually significant. The economic implication (or political implication) is that the offsetting forces in the voting are not important, but it is the "residual value" that determines the outcome. For example, suppose there are 1 million eligible voters in a country, of which 300,000 of them vote to the left and 400,000 vote to the right, so only 100,000 votes actually matter. According to the postulate of self-interest, in democracy, politicians pursue vote maximization. The median-voter theorem implies that what a politician should do is to get the votes that can overwhelm 100,000 (instead of 300,000 or

400,000) with the least cost. Thus, it is the interests of the minority instead of the majority) that are important.

By contrast, in the market, any interests will be cared about, whether of the majority or the minority. Just as mentioned in Lecture 9, big chickens eat big rice, small chickens eat small rice, and even broken rice is eaten by worms. As long as there is demand, there will be always supply and there will be always a market to take care of the demand.

Section 5: Democratic Voting Infringes on PPRs

According to the previous analysis about democracy, there are problems as follows:

(1) Democracy will harm the interests of the minority, which may not be lower than those of the majority. For example, one public choice may benefit 10 persons and harm only one, but the beneficiaries only get the benefits of 10 units, while the harmed suffer a loss of 1000 units.

(2) The representatives in indirect democracy vote for the interests of their own instead of all. What is more, the outcome of voting may not truly reflect the interests of the majority.

(3), The most serious problem of democracy lies in that if there is no appropriate and clear division between private space and public space, and the matters should have been private decision are wrongly regarded as public choice, democratic voting will infringe on PPRs rather than just the interests of the minority. Because everyone will inevitably be a minority in some cases, the infringement on PPRs of the minority by democratic voting will be eventually expanded to infringing on PPRs of all!

There is an imaginary example: If there were a voting to decide whether the properties of the richest man should be divided up, could it be passed? The answer is yes. People are self-interested, so if they could get money by voting, why not agree? They would make some grandiose excuses such as accusing the richest man of being nothing more than an unscrupulous bloodsucker parasitic on the public to conceal their real desires.

In developing countries, the essence of nationalization and land reform (taking lands away from landlords to be divided up among tenants) is precisely such an imaginary example in reality. In developed countries, there seem to be no such voting of openly robbing the rich, but the essence of "evil policies" including price control, minimum wage law, trade union, quantity control (such as purchase limitation and quota), social endowment insurance, trade protectionism and in fact all policies with purpose of income redistribution is also precisely such an imaginary example in reality. The essence of income redistributing is an infringement on the right to income right, and because such a policy often only redistributes income, but does not re-delineates PPRs, it will cause rent dissipation. In developed countries, all the policies of income redistribution are approved by democratic voting.

The appropriate and clear division between private space and public space is precisely the delineation of property rights. In other words, if democracy is implemented before the delineation of property rights, self-interested people will make use of voting to rob others' private properties. It is the main cause of "tyranny of the majority", and the harm to the interests of the minority is just part of it.

In modern times, the most painful lesson caused by democracy is the bloody French Revolution. The political scientists generally believe that it is due to the problem of direct democracy. However, the real cause is democracy without delineation of rights. It was in the National Assembly where the murders were approved. People always demonstrated on the streets, which was direct democracy and put tremendous pressure on the National Assembly, but after all, the decisions to send one after another to the guillotine were not the outcomes of referendum. It was unrestrained democracy to decide whether to deprive one of his or her life by voting in parliament, rather than dictatorial judgments according to the laws (contractual terms delineating rights) in court. In fact, it is far more than this time in history to kill people by democratic voting. As all know, Socrates, one of the greatest philosophers in ancient Greece, was sentenced to death

by Athenians through democratic voting. The French Revolution was more dramatic in that Robespierre, who had approved many murders in parliament, was finally voted to be sent to the guillotine too. As mentioned just now, democratic voting that infringes on the interests of the minority will eventually infringe on those of all, because everyone will inevitably be a minority in some cases, which is tested by the blood of Robespierre.

Since ancient times, in the western political thinking, the reputation of democracy is notorious. In his "Politics", Aristotle recorded and evaluated the political systems of various ancient Greek city-states he observed. He thought that three systems were good and three were bad, among which democracy was one of the bad systems.

Democracy changed to a good system when the US was established. Strictly speaking, it is American democracy that is regarded as a good system. What is American democracy? The essence of it is constitutional democracy, which is a democracy where voting is restrained by constitution known as the Bill of Rights or the first to the tenth amendments of the US Constitution. Bill of rights mainly prohibits Congress from legislating on certain matters, which means democratic voting is prohibited in these designated areas! However, on the one hand, the laws written on paper are not the laws really implemented in reality. The constitutions of many developing countries (especially South American countries) almost copy the Constitution of the US word by word, but most of them do not really implement constitutionalism. On the other hand, even the Constitution of US is ambiguous, which causes the American democracy degenerates to a bad institution eventually. With the Bill of Rights to prohibit the legislation of Congress (democratic voting) from infringing on PPRs, it is impossible to hold a referendum, or a voting in Congress to murder people as during the French Revolution, or openly divide up the properties of the richest man, but the legislations with the purpose of income redistribution seem not to infringe on PPRs, so they can be approved in Congress.

If the use of democratic voting is not clearly and strictly limited to the affairs in public space, combined with the fact that representatives vote for the interests of their own rather than of all, and the median-voter theorem, there must be interest groups in the voting system. Interest group refers to a small group of people unite for common interests and take common actions in voting. Interest group uses votes as political resources to harm others, because they are committed to preventing changes that are beneficial to society but unfavorable to this small group of people. In the democratic countries, it is impossible to push forward major reforms, because people with vested interests will kill it early by voting!

READING MATERIAL: HOW CAN INTEREST GROUPS AND DEMOCRATIC VOTING LEAD TO "TYRANNY OF MINORITY" AND EVEN "ALL LOSE"? ¹

In addition to the traditional problem of "tyranny of the majority" in democracy, a small group of vested interests unites to form an interest group to influence the voting outcome, which will even lead to "tyranny of the minority". Next we will show how it happens in democracy with the example of protectionism in international trade.

Under the institution of democratic election by one-man-one-vote, the strength of single individual is too weak. However, self-interested people's intention to influence the voting outcome will not be weakened. The solution is simple: unity is strength! Those with common interests can unite to vote. Thus, the institution of democratic election by one-man-one-vote will inevitably give birth to interest groups.

An interest group must be formed by a small group of people, because the fewer people, the easier it is for them to have common interests, and the easier it is to reach consensus and take common actions. After all, different people still have different interests, so obviously, the fewer people, the easier it is for them to have common interests, which implies there are transaction cost for those with different interests agree with each other.

Ceteris paribus, the transaction cost of consensus increases with more people. For example, it is easy for dozens of students in a class to reach an agreement on where to travel, but it is very difficult for thousands of students in a school to do so.

Since the fewer people, the easier it is to form an interest group with common interests, and the easier it is to reach an agreement and take common actions in voting, the interests represented by an interest group can only be those of a small group of people instead of the public. Of course, if there are too few people in an interest group, there will be too few votes to influence the voting outcome, but in any case, it is better to unite to vote that insures their voting will not be offset.

How will politicians choose in democracy? Will they choose to represent the interests of the majority because democratic voting claims to use the minority rule? It is a naïve idea to think so. Politicians are self-interested, too. Under the specific constraints of democratic voting, politician will try to maximize the votes and minimize the cost of canvassing. With the constraint of interest groups, politicians will not foolishly make solicitation of the votes one by one, but the votes of interest groups. As long as a politician has persuaded the leader of an interest group, the leader can order all the members in the interest group to vote for him.

Next, we will demonstrate the terrible consequence of democracy with a numerical example.

As supposed in the example when the median-voter theorem was introduced, there are 1 million eligible voters in a country, of which 300,000 vote to the left and 400,000 vote to the right, so only 100,000 votes actually matter. Will a politician make solicitation of the 100,000 votes one by one? Of course not. He will try to win the votes of interest groups. Suppose there are 10,000 voters in an interest group on average, which means as long as the politician has persuaded the leader of an interest group, he can win 10,000 votes. In theory, he only needs to persuade 11 leaders of interest groups to win 110,000 votes.

Thus, can voting really be good at least to the majority? With interest groups, the united minority can defeat the disunited majority.

How to persuade the leaders of interest groups? Attention: we are not talking about bribery, black money politics, or vote buying that break the laws about voting. We are talking about the completely legal and known-to-all lobbying. The term of "lobby" originally meant a large entrance or reception room or area, but later was extended to refer to an interest group who tries actively to influence legislation in their favor in exchange for their political resources. What political resources do they have? Votes, of course! If a politician promises to help them approve the bills in their favor, the leader will order the members to vote for him. The politician does not need to care the interests of the public, but just needs to satisfy the interest groups. Thus, in democracy of reality, the interests of the majority actually do not matter.

However, the minority should not be triumphant to think they can be benefited by harming the interests of the majority. The horror of democracy not only lies in the tyranny of minority, but also the fact that it will ultimately harm all. Think more deeply of the above example. The politician must get enough votes from 11 interest groups. For example, there are interest groups in agriculture, textile industry, steel industry, and so on. Well, he has had enough votes, and is successfully elected. He is quite creditable, so he takes seriously all the promises he made in the election campaign and really fulfills them one by one. Therefore, he puts forward and makes sure 11 bills to be approved in parliament, each of which is beneficial to one of the interest groups at the expense of all the others in the country.

For example, he tries to approve an act of agriculture for the agricultural group that would implement trade policy of protectionism in agriculture and forbid foreign products to be imported to the country. As a result, the prices of agricultural products are much higher than before, so those engaged in agriculture benefit from the much higher income. However, the consumers have to pay higher prices to buy agricultural products, so they are all harmed.

Similarly, the politician tries to approve an act of textile for the interest group in textile industry, and those in textile industry benefit, but the consumers have to pay higher prices to buy clothes, shoes, socks, etc. and are harmed.

The key is: who are the consumers of textiles? Some of them are precisely those engaged in agriculture. And who are the consumers of agricultural products? Some of them are precisely those engaged in textile industry.

One is engaged in the production of one industry, but he or she is the consumer in all the other industries at the same time! If only the industry he is engaged in is protected, he can benefit without harm. However, the politician cannot just make solicitation of the votes in one interest group, because they are not enough for him to be elected. When the other industries are also protected, he will suffer as the consumer of the other products. The politician exchanges the legislations in favor to the interest groups for their supports in voting. Each bill approved in parliament is beneficial to one interest group at the expense of all the others, so there will be no beneficiaries eventually in democracy.

Some may ask since the democratic voting will attract interest groups, which will harm all, can the establishment of interest group be forbidden? However, it is impossible to prohibit people from uniting together to express their interests in a free society. According to the postulate of self-interest, it is impossible to suppress people's nature to pursue their own interests. How about if people are informed that interest groups will eventually harm all? If there were only one interest group, the member in it can benefit at the expense of the others, so everyone will unite to protect his own interests as soon as possible, otherwise he will be harmed. In other words, the problem lies not in interest groups, but in democratic voting. Interest groups will harm all, which is only an inevitable result under the constraint of democratic voting.

In a word, with the human nature of self-interest, politicians will sacrifice the interests of the country in exchange for the votes of interest groups, but these united minorities are also the majority as consumers in the other industries and cannot be free from infringements on their rights. Thus, with interest groups, democracy has changed from an institution based on the majority rule that where the minority harm the majority, and eventually harm all. Is democracy really desirable?

1. This reading material is from another book of mine "The Truth of Trade", P147, China CITIC Press, March, 2015. I have made some modifications for this book [2](#)

Section 6: The Difficulties of Dictatorship

There are also difficulties for dictatorship that seem to be two kinds.

(1) Representativeness. In dictatorship one represents all, while in (indirect) democracy, a few (the parliamentarians) represent all. However, the key to a good representation lies not in the number of delegates, but in whether the interests of them are closely related to those of all. In the example of a rented house, the owner can represent the interests of multiple tenants in use of the public space in the house better than the committee by election, because his income comes directly from the tenants, and he is also under the competitive pressure of other houses.

(2) Arbitrariness.

These two difficulties can actually be regarded as one: the dictator may not be wise enough. If a dictator is wise, he will recognize his own interests are closely related to the interests of all. Especially for the dictatorship of Monarchy, just as the property rights of the whole house (including part of the house that is public space for tenants) is clearly delineated to the owner, the property rights of the whole country is clearly delineated to the dictator. There are public spaces for specific users, but they are all private properties of the dictator, so there is no reason for a wise dictator not to take care of the overall interests of all. What is more, a wise dictator knows well that there is information cost for him to know everything, so he will try his best to appoint talents and abilities as his ears and eyes to reduce the blindness of arbitrary decisions as much as possible. A perfect dictator needs to be wise instead of kind. A wise man will never be cruel because cruelty is no good for his interests, but silly kindness will lead to doing harm with good intentions. Most poor-performed emperors in ancient China were by no means cruel but fatuous in politics. In fact, the great thinkers in both the East and the West in ancient time all advocated that the most ideal state is the dictatorship of the wise monarch (the East) or the philosophy king (the West).

The real difficulty of dictatorship is that if the dictator is unwise, theoretically, power is concentrated in his hands, so it is difficult to get rid of him. Palace coups and uprising will cause bloodshed, but the transaction cost may be not higher than that of the election campaign every two years for parliament and every four years for presidency in modern western countries. The transaction cost of a court coup should be the lowest, because it involves a much narrower scope and generally a much lower frequency than the regular election. Uprising will lead to war, so the transaction cost may be the highest. However, not only the developed countries should be focused on, but also the developing countries should be taken into account. Most of them are democratic countries, and there are political turmoil and even wars from time to time. In dictatorship, the dictator only needs to be wise, while in democracy, it is not enough that the politicians are wise if there is no constitution to restrict voting with appropriate boundaries.

There is a problem similar to this difficulty of dictatorship: how to choose a wise successor. The hereditary system has a great advantage in delineating rights and minimizing the transaction cost of power struggle, which is precisely why it is adopted in ancient societies. However, there is no guarantee that the wisdom can also be inherited like blood lineage. Education has some use, but it is far from 100% reliable. It is not only a political problem, but also family businesses are all facing the same difficulty in inheritance.

There has been Japanese who translated a famous Chinese ancient book about "Imperial Art" into Japanese, and the preface for the Japanese edition was thought-provoking. He said, some may wonder, modern society is democratic, and what is the value in reading such an ancient book on imperial art? He thought this book was actually more valuable in modern times than in ancient times! Because in ancient times, only an emperor needed to study imperial art, but now every boss is a little emperor in his enterprise and needs to master it. Obviously, this translator has a good understanding about the essence of dictatorship, and knows that dictatorship with a wise dictator is the best institution, but the difficulty lies in how the dictator can be wise, and one of the means is learning from the ancient monarch. Thus, even if there is not monarchy in politics in modern society, it is still of great significance to study it deeply to solve the problems in enterprises, especially family enterprises.

As for the famous saying of Lord John Acton, "Power corrupts; absolute power corrupts absolutely" which is often used to criticize dictatorship, it can actually not hold up to closer inspection. What does "power" mean? Does it mean PPRs? What does "absolute power" mean? Does it mean clearly delineation of property rights? How can it corrupt the owner? What does "corruption" mean? Usually it means rent seeking in economics that is caused by administrative monopoly, precisely an infringement on PPRs. This statement may mean that one with huge power will easily become headstrong or unwise. However, that one is unwise is not necessarily related to he has huge power for a long time. More probably, his decision-makings have been proved to be correct, so he becomes too conceited and cannot accept others' different opinions. During the Chinese history, Kangxi of Qing Dynasty, who was known as one of the best emperors, reigned for 60 years, while Emperor Yang of Sui Dynasty, who was known as one of the worst emperors, reigned for only 14 years. It is not power that corrupts people, but the stupid ones who have huge power can do great harm. It is not absolute power that absolutely corrupts people, but if absolute power falls into the hands of the stupid ones, they can only be stopped by uprisings with very high transaction cost.

Section 7: Three Utopias

In MSE, it has been proved that, an omniscient and selfless dictator, who arranges production and consumption according to everyone's preference and capacity, can have the same result (Pareto-optimality) as that everyone produces and consumes according to the guidance of price in the market. MSE realizes that the dictator must be completely wise (having all the information and knowing that the overall interests of the society are his own interests), but it does not realize that the transaction cost of using the market must be zero.

With the addition of the analysis in this lecture, if direct democracy is used and everything is approved by unanimous voting, the result will be the same as that of the market with zero transaction cost or dictatorship with complete wisdom. Since it is direct democracy, all have participated in the decision-making, and since everything is approved by unanimous voting, no one's interests will be harmed, which must be Pareto-optimal.

As mentioned in section 1, there are three ways to make decisions (choices) on different affairs: market, dictatorship and democracy. Corresponding to these three ways, there are three utopias: the market with zero transaction cost, the dictatorship with complete wisdom, and the direct democracy with everything approved by unanimous voting. They are all utopians because the results of them are all Pareto-optimal, but they are all impossible in reality.

It is impossible to use the market without transaction cost, and there will be no market when the transaction cost of it is too high relative to the benefit of it. It is not easy to delineate the property rights of public space. Although they can be delineated and distributed in the form of shares in theory, if there are too many people involved (there are hundreds or thousands of owners in a community, not to mention the eligible voters in a country), the transaction cost of negotiation will be very high. Due to the information cost (a kind of transaction cost), no dictator can be completely wise. Also due to the information cost, it is impossible for all in a society to have a deep understanding of all affairs and know how to vote for their own interests. And the cost of negotiating with every voter and modifying the proposal to satisfy everyone is as high as that of negotiating in a market transaction. In other words, it is transaction cost that determines there are three different ways (market, dictatorship and democracy) to make decisions, and it is also transaction cost that determines there are three corresponding three utopias!

The final conclusion: political economics is a difficult field, not because the theories are different from those in other fields of economics, but because studiers need to try their best to get rid of the interference of subjective feelings (such as ideology) on objective analysis. Nowadays, too many take democracy for good and dictatorship for bad without thinking. They completely ignoring that the political institutions or the modes of public choice in most times of human society is dictatorship, and also completely ignores that even in modern times, people seldom make decisions or choices by voting except in politics.