

WWS 507c: Quantitative Analysis

Lecture 1a: What is Statistics?

Eduardo Morales

Princeton University

September 11, 2014

What is Statistics?

- Statistics helps us process data or information in order to answer questions such as
 - Why does the casino make a profit at the roulette?
 - Why insurance companies do not go bankrupt?
 - Does Nadal serve better than Djokovic?
 - Has student A better grades than student B?
 - How many people are employed? unemployed?
 - What percentage of Americans support the death penalty?
 - What can you use to predict first-year college grades?
 - How does Netflix know what kind of movies you like?
 - Does smoking cigarettes cause cancer?
 - What are the effects of new medical treatments?
- We can group these questions into five categories...

Probability

- Category 1. Theory of Probability.
 - It helps answer questions such as:
 - Why does the casino make a profit at roulette?
 - Why insurance companies do not go bankrupt?
 - It focuses on assessing risk and other probability-related events.
 - Much statistical reasoning depends on the theory of probability. The connexion is through chance models or models of randomness. These models impose assumptions on outcomes of uncertain events (e.g. probability of obtaining an even number when rolling a die).
 - Besides, very large industries exist thanks to the theory of probability:
 - Gambling industry.
 - Insurance industry.
 - Data forensic firms that identify patterns that suggest cheating in an exam.
 - Large risks arise from wrong usage of the theory of probability:

Probability

- Probability can make you win money!
- Monty Hall problem.
- On the TV show *Let's Make a Deal*, participants are offered a choice of one of three doors.
- Behind one of the three doors, there is a grand prize, and behind the other two there is nothing.
- After you pick the door, the host, Monty Hall, who knows which door hides the big prize, always opens a door that is not the one containing the big prize nor the one that you initially pick.
- He shows you the empty door and then asks you if you want to switch doors.
- Shall you change your choice?

Probability

- Marilyn vos Savant, listed in the Guinness Book of World Records Hall of Fame for “Highest IQ”, set off a nationwide furor when she discussed the Monty Hall problem in her syndicated column *Ask Marilyn*.
- She gave the correct answer and then received more than a thousand letters, many from college professors, most of them saying that she was wrong. A math professor literally wrote:

You blew it! As a professional mathematician, I'm very concerned with the general public's lack of mathematical skills. Please, help by confessing your error and, in the future, being more careful.

- Marilyn stuck to her answer and invited people to play the game at home. Computer simulations were run at the Los Alamos National Laboratory in New Mexico. Marilyn was right.

Probability

- Another probability paradox involves a man named Smith who is walking with his daughter and says that he has another child at home.
- Taking into account that the probability of having a boy or a girl is exactly 0.5 and the gender of the first child does not affect the gender of the second child, what is the probability that the child at home is also a girl?
- Now imagine that Smith had told you that the daughter he is walking with is actually the oldest one of his two children, what is the probability that the child at home is also a girl?
- If Smith was equally likely to be walking with this oldest or his youngest children, would the probability that the child at home is also a girl change?

Probability

- Is it possible that Lance Armstrong was innocent?
- Imagine a drug test that has 95% of accuracy.
 - An athlete that has consumed drugs will be tested positive in 95% of the cases.
 - An athlete that has not consumed drugs will be tested negative in 95% of the cases.
- Imagine that only 5% of the athletes really consume drugs.
- Given that Lance Armstrong was tested positive, what is the probability that he actually consumed drugs?
- In order to simplify the exercise, imagine that 10,000 athletes are tested each year.

What is Statistics?

- Category 2. Descriptive Statistics.
 - It helps answer questions such as:
 - Does Nadal serve better than Djokovic?
 - Has student A better grades than student B?
 - The “percentage of 1st serve points won” or the GPA of a student are descriptive statistics. Studies typically produce so many numbers that summaries are needed.
 - Descriptive statistics are easy to calculate, easy to understand, and easy to compare across samples or individuals. However, they are not perfect. The “percentage of 1st serve points” won does not reflect the skills of the players receiving the serve. The GPA does not reflect the difficulty of the courses that different students may have taken.
 - Btw, according to ATP stats, both N and D win 77% of points with their 1st serve. D is better than N in percentage of 2nd serves won (59% vs. 57%). However, N has a higher percentage of service games won (89% to 87%)....

Q: Who is better at serving? A: i?i?i?i?i?i?i?

Descriptive statistics can be misleading!

- Wellfleet is a small Massachusetts town known for its oysters and artists.
- There was considerable surprise when a Boston newspaper reported that Wellfleet had the highest murder rate in Massachusetts in 2005, with 40 murders per 100,000 residents that year.
- Boston only had 17 murders per 100,000 residents.
- A puzzled reporter looked into this statistical murder mystery and found that no Wellfleet police officer could remember a murder ever occurring in Wellfleet.
- However, a man accused of murdering someone twenty miles away had turned himself in at the Wellfleet police station, and this Wellfleet arrest had been erroneously annotated as a Wellfleet murder.
- Because Wellfleet had only 2,491 residents, this one misrecorded arrest translated into 40 murders per 100,000 residents.

Descriptive statistics can be misleading!

- Boston, in contrast, had 98 murders, which is 17 murders per 100,000 residents.
- This murder mystery shows how randomness and recording errors can make a big difference if the base is small.
- A misrecorded murder in Boston has little effect on its murder rate.
- A misrecorded murder in Wellfleet puts a small village known for its oysters on par with Detroit.
- One way to deal with a small base is to use data for several years to get a bigger base.

What is Statistics?

- Category 3. Inference.
 - It helps answer questions such as:
 - How many people are employed? unemployed?
 - What percentage of Americans support the death penalty?
 - In general, it is expensive and difficult to count the employed population in a large geographical area or to ask a question to every American citizen. A sample is a subset of the population of interest (e.g. American citizens).
 - Statistical inference allows to do valid generalizations from samples. It indicates which assumptions are needed so that we can use information from a sample to say something about a population.
 - A political poll is one form of sampling. The polling firm Gallup reckons that a poll of 1,000 randomly selected households will produce roughly the same results as a poll that attempted to contact every household in America.
 - We will learn how to extract information from samples: if 53% of a sample of 1000 individuals support the death penalty, with which degree of confidence can you conclude that a majority of Americans support the death penalty?

What is Statistics?

- Category 4. Prediction.
 - It helps answer questions such as:
 - What can you use to predict first-year college grades?
 - How does Netflix know what kind of movies you like?
 - Prediction is based on correlation. Correlation measures the degree to which two phenomena are related to one another. The correlation between SAT composite score and first-year college GPA is 0.56 (computed by The College Board). Therefore, SAT grades are good predictors of college grades.
 - Netflix knows how you rated the movies that you already watched. Netflix compares your ratings with those of other customers to identify whose ratings are highly correlated with yours. Netflix recommends films that like-minded customers have rated highly but that you have not seen.
 - Correlation does not imply causation: a positive or negative association between two variables does not necessarily mean that a change in one of the variables is causing the change in the other. If a grading mistake assigns you an unfairly high SAT score, this will not affect your college grades.

Regression Toward the Mean

- Horace Secrist was a professor of economics at Northwestern University.
- He studied business practices. Secrist and his assistants spent 10 years collecting and analyzing data for 73 different industries in the United States, including department stores, clothing stores, hardware stores, and banks.
- He compiled annual data for the years 1920 to 1930 on several metrics of business success, including the ratio of profits to sales.
- For this ratio, he divided the companies in an industry into quartiles based on the 1920 values of the ratio: the top 25 percent, the second 25 percent, the third 25 percent, and the bottom 25 percent.
- He then calculated the average ratio for the top-quartile every year from 1920 to 1930. He did the same for the other quartiles.
- In every case, the quartile ratios converged over time. The companies in the top 2 quartiles in 1920 were more near the average in 1930. The companies in the bottom 2 quartiles in 1920 were more near the average in 1930. The most extreme quartiles showed the greatest movement towards the mean.

Regression Toward the Mean

- Secrist had discovered a universal economy truth. Over time, the most successful and least successful firms tended to become more near the average.
- Secrist then went on to write a book that was called *The Triumph of Mediocrity in Business*. In this book, Sacrist explained through complicated economy theories why industry was evolving towards a world in which all firms will be identical to each other.
- A similar conclusion was reached by Sir Francis Galton when comparing the heights of parents and their adult children. Unusually tall parents tend to have somewhat shorter children, while the reverse is true for unusually short parents. The erroneous conclusion is that heights are regressing towards mediocrity; indeed, Galton titled his study *Regression Towards Mediocrity in Hereditary Stature*.
- Is it surprising that we have not reached yet a world in which all firms have identical ratios of profits to sales and all humans are of the same height?

Regression Toward the Mean

- Let's work with a detailed example. Suppose that there are three kind of firms in an industry:
 - strong: each strong firm has an average profit to sales ratio of 40 percent
 - middle: each middle firm has an average profit to sales ratio of 30 percent
 - weak: each weak firm has an average profit to sales ratio of 20 percent
- A firm's average profit is its "ability". Each company's profit in any year is equally likely to be its ability plus 6, plus 3, minus 3, or minus 6.
- Thus, a firm with an ability of 40 is equally likely to have a profit of 46, 43, 37 or 34 percent in any given year.
- Their actual profits fluctuates randomly around their ability, which is constant.
- Use these assumptions to generate data for 2 years, which we can call 1920 and 1930.

Regression Toward the Mean

- There are 12 equally likely observed profits percentages in each year: 46, 43, 37, 34, 33, 27, 26, 24, 23, 17 and 14.
- Following Secrist, suppose that we group firms into four groups of equal size based on their observed profit in 1920. E.g. the top quartile includes firms with observed profits of 46, 43, and 37.
- Let's call these four groups as: "1st quartile, 1920", "2nd quartile, 1920", "3rd quartile, 1920", and "4th quartile, 1920".
- Compute the average profit in 1930 of those firms included in the group "1st quartile, 1920". Is it larger or smaller than their average profit in 1920?
- Similarly, compute the average profit in 1930 for those firms included in the group "4th quartile, 1920". Is it larger or smaller than their average profit in 1920?

Regression Toward the Mean

- Understanding regression towards the mean is very useful.
- On one occasion, a senior instructor of the Israeli army said

On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So I have concluded that punishment and screaming works better than reinforcement.

- Is there any evidence that punishment actually helps trainees progress faster than praise?
- What do you think about sports journalists that claim that teams that switch coaches in the middle of the season tend to win the first game after the switch because players try to impress their new coach and this makes them perform better?

What is Statistics?

- Category 5. Identifying causal relationships.
 - It helps answer questions such as:
 - Does smoking cigarettes cause cancer?
 - What are the effects of new medical treatments?
 - In order to identify a causal relationship, we would like to see each individual in a population both smoking and not smoking/taking and not taking the new medical treatment. This is obviously impossible, so statisticians *define* causal relationships in terms of randomized controlled trials (RCT).
 - In a RCT, individuals in a sample are randomly assigned to different groups. These groups receive different treatments. There is a causal relationship between a treatment variable and an outcome variable if these different groups obtain (on average) different outcomes (e.g. double-blind clinical trials).
 - In many cases, RCT are not possible (e.g. effect of education on wages). In this case, statistics uses regression analysis to obtain information about causal relationships from observational data.

Causes of Cholera

- In some cases, policy changes generate a variation in the data that is similar to the variation that would arise in an ideal RCT. This was the case with cholera.
- Cholera hit London in 1832 and killed 6,500 people.
- The medical establishment at the time believed that cholera was caused by breathing “miasma”, or noxious air.
- The miasma theory was supported by the fact that cholera was more commonplace in poor neighborhoods that lacked proper sanitation services and smelled awful.
- The miasma theory could not be proven definitely because other confounding factors needed to be taken into account.
- For instance, people living in poor neighborhoods tended to be older than in other neighborhoods, ate different foods, had different occupations, and had little or no heat.

Causes of Cholera

- Doctors could not implement a RCT that allowed to test the theory that cholera was caused by miasma. They could not force randomly selected people to breath poisonous air while other selected people lived in a clean-air environment.
- In 1848, London connected all the houses to sewer lines. These sewer lines deposited the raw sewage into the Thames River which, directly or indirectly, was the source of drinking water for many Londoners.
- The 1848-49 cholera epidemic quickly followed the construction of the sewage system.
- This made the doctor John Snow hypothesize that cholera was caused by drinking water which had become polluted by sewage.
- Using this fact to infer that polluted water causes cholera would have been an example of a logical fallacy: the fact that one event happened shortly after another does not mean that the former necessarily caused the latter.

Causes of Cholera

- For several years, two water companies, the Southwark and Vauxhall Water Company and the Lambeth Company, supplied water to the same London neighborhoods through different pipes using water drawn from the same polluted part of the Thames River.
- In the 1848-49 cholera epidemic, their customers had similar death rates.
- The Water Act of 1852 mandated that London water companies stop taking water from the heavily polluted part of the Thames River by August 31, 1855.
- Lambeth actually made the move in 1852, so that it started to draw water from the Thames before it became polluted by the London sewage.
- Southwark and Vauxhall did not relocate until 1855.
- Snow realized that this was a perfect opportunity to test his theory.
- Snow examined all the recorded cholera deaths during 1854 and determined which homes received their water from each of the two companies.
- Snow found that the rate of death per 10,000 customers was nearly nine times higher for Southwark and Vauxhall customers

What is Statistics?

- In WWS 507c:
 - we will learn a little bit of
 - Probability theory.
 - Descriptive statistics.
 - Prediction theory.
 - and emphasis will be given to the study of
 - Inference.
 - Causal analysis.

WWS 507c: Quantitative Analysis

Lecture 2: Probability

Eduardo Morales

Princeton University

September 16, 2014

INTRODUCTION

Definition of random phenomenon

- Let's flip a coin.
- Can you tell what the outcome will be?
- If we were to flip the coin many many times, would you be able to tell the proportion of times that the outcome of the coin flip will be heads? (assume that the coin is fair)
- If the answer to the first question is “no” and the answer to the second one is “yes”, then you have just defined the concept of a random phenomenon.
- A **random phenomenon** is a phenomenon such that
 - individual outcomes are uncertain; and
 - the proportion of times with which each potential outcome will be observed in a **large number of independent repetitions** is certain.
- The **probability** of any outcome of a random phenomenon is the proportion of times the outcome occurs in a very long series of independent repetitions.

Law of Small Numbers

- Probability only describes what happens **in the long run**.
- Many people **wrongly** infer probabilities from regularities that appear in short sequences of outcomes of random phenomena.
- For example, a few years ago a man won the Spanish national lottery with a ticket that ended in the number 48. Proud of his accomplishment, he revealed his theory: *I dreamed of the number 7 for seven straight nights, and 7 times 7 is 48!*
- In the 2010 NBA finals between the Boston Celtics and the LA Lakers, Celtic guard Ray Allen made seven three-point shots in a row. This is an example of what journalists call “hot streak”. Observers usually tend to theorize about the reasons underlying hot streaks and cold streaks. They also tend to be surprised about the existence of these hot and cold streaks.

Law of Small Numbers

- For example, one of Allen's teammates said it was "incredible". Another called it "unbelievable". One sportswriter wrote that "Allen had slipped into that shooting zone only visited by real-life superstars and movie characters".
- If they had known a little bit more about probability, what they should have said is: "Allen has taken more than 7,000 three-point shots in his career and made 40 percent of them. It is, in fact, almost certain that there will be streaks of seven successful shots in a row at some point in those 7,000 tries. We just got lucky that it happened during the NBA finals."
- Daniel Kahneman and Amos Tversky worked on identifying ways in which our judgement is affected by systematic biases and errors. One of the cognitive errors Kahneman and Tversky observed is belief in the law of small numbers.
- The belief in the law of small numbers makes people extract knowledge about underlying probabilities after observing very few outcomes from a random phenomenon.

Law of Small Numbers

- Let's imagine that we extract 5 balls from a container. We do not know how many red and blue balls are in the container. If we draw five balls and four are red, we reason (incorrectly) that 80 percent of the balls in the container must be red. So, there is an 80 percent chance that the next ball will be red, too.
- These misconceptions cost a lot of money to certain organizations.
 - imagine the container is “James Rodríguez” and the two types of balls are “soccer genius” and “not soccer genius”,
 - our five balls are the five games he played during the 2014 World Cup in Brazil,
 - the outcome of the five random draws were actually “soccer genius”,
 - Real Madrid wrongly inferred that the probability of “soccer genius” was one and paid 103 million dollars for his transfer,
 - sadly for Real Madrid fans, the next four draws have been “not genius”.

Randomness

- What does randomness really mean?
- Deterministic physical laws govern what happens in the flip of a coin.
- The uncertainty in the velocity with which we flip the coin is key for the randomness of the phenomenon.
- A coin toss is basically deterministic. The coin obeys Newton's laws of motion, with its final state depending on the velocity with which we flip the coin, its rate of spin, and time traveled.
- For tosses where the coin spins rapidly and goes high in the air, the set of initial velocity values that lead to either heads or tails are of equal size.
- So uncertainty about the velocity with which we flip the coin makes the phenomenon random.
- The same is true for random numbers generated by computers. Computers follow deterministic algorithms. It is our lack of knowledge about the initial state of the algorithm what makes the generated numbers random.

Probability Theory vs. Statistical Inference

- In this lecture, we assume that we know the prob. of random phenomena.
 - e.g. we know that the probability of observing heads in a coin flip is 0.5.
- Given the knowledge of probabilities, in this lecture, we will learn what kind of outcomes we are likely to see when we observe random phenomena.
 - e.g. what is the probability of observing HHTT if we flip a coin 4 times?
- In future lectures, we look at the situation from the opposite point of view: given that we have observed certain outcomes from a random phenomenon, what can we say about the probabilities that generated these outcomes?
 - e.g. given that I flipped a coin 4 times and I obtained HHTT, which is the best possible guess for the probability of obtaining heads when flipping that coin?

The Most Important Random Phenomenon

- In this lecture, to gain intuition, we will study multiple different random phenomena: flipping coins, throwing dice, etc.
- In future lectures, we will focus all our attention on a very special type of random phenomenon:

OBTAINING A RANDOM SAMPLE FROM A POPULATION OF INTEREST

- E.g. if the population of interest is the set of American citizens with voting rights, an outcome of this random phenomenon would be the specific set of people polled by Gallup.
- You can think of this random phenomenon as the process of randomly extracting 10 balls with replacement from an urn with a 100 balls. The outcome of this random phenomenon is the specific set of 10 balls we extract.

PROBABILITY MODELS

Probability Model

- Given the difficulty to solve questions that relate to random phenomena relying exclusively on intuition, we resort to the help from mathematics.
- A description of a random phenomenon in the language of mathematics is called a **probability model**.
- A probability model tells us everything that we may know about a random phenomenon before we observe its outcome.
- A probability model tells us:
 - the set of all possible outcomes of a random phenomenon (this is the **sample space**)
 - the probability of each of the elements in the sample space
- For example, for the random phenomenon of “rolling a die”, the sample space is

$$\{1, 2, 3, 4, 5, 6\}$$

and the probability of each element of this sample space is $1/6$.

Probability Model

- The name of *sample space* makes reference to the specific random phenomenon of “obtaining a random sample from a population of interest” (see slide 9). In this case, each possible outcome is a sample from the population of interest. The sample space contains all possible samples.
- We will also use the term **event** to refer to an outcome or a set of outcomes of a random phenomenon. That is, an event is a subset of the sample space.
- E.g. think of the random phenomenon “flipping a fair coin twice”. The sample space is:

$$\{HH, HT, TH, TT\}.$$

The probability of each outcome is 1/4. We can describe an event as “obtaining at least one head”. The set of outcomes included in this event is

$$\{HH, HT, TH\}.$$

ASSIGNING PROBABILITIES TO OUTCOMES

Assigning Probabilities to Outcomes

- How do we know that the probability of the outcome HH in the random phenomenon “flipping a fair coin twice” is $1/4$?
- As we saw before, the probability of an event is the percentage of times that this event will happen, when the experiment is repeated
 - ➊ over and over again,
 - ➋ independently, and
 - ➌ under the same conditions.
- Therefore, assigning probabilities to outcomes often requires long observation of the random phenomenon.
- In some specific cases, we are willing to assume that all outcomes in a sample space are equally likely because of some knowledge of the physical process generating such outcomes. As we saw before, the physics behind coin tossing reveal that fair coins have a physical balance that should make heads and tails equally likely.

Assigning Probabilities to Outcomes

- If all outcomes in the sample space are equally likely, then, for outcome A ,

$$P(A) = \frac{1}{\text{number of outcomes in } S},$$

where S is the sample space.

- Therefore, for random experiments in which it is simple to compute the total number of outcomes (i.e. total number of elements in the sample space) and they are all equally likely, computing the probability of each of these outcomes is very simple.
- Example: die rolling with $A = \{1\}$. The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. If the die is fair (i.e. equal probability of each outcome in the sample space), then $P(A) = 1/6$.

Assigning Probabilities to Outcomes

- In some random phenomena, computing the total number of outcomes in the sample space can be hard. Luckily, mathematicians have come up with convenient formulas. These are called *counting formulas*.
- Here we will learn three different counting formulas through three exercises.

Assigning Probabilities to Outcomes

- **Exercise 1:** balls numbered 1-100 are placed in a box. Four balls are chosen at random, 1 by 1 and *without replacement*. What is the probability of the outcome (53, 67, 12, 99)?
- An elementary event or outcome is an *ordered* sequence of 4 numbers.
- We need to count the number of possible ordered sequences of 4 numbers that can be drawn from a set of 100 numbers *without replacement*.
- **Counting formula 1:** The number of different ordered sequences of size k that can be drawn from n objects without replacement is

$$n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

- Therefore, $P(A = (53, 67, 12, 99)) = \frac{1}{100 \times 99 \times 98 \times 97} = \frac{1}{94,109,400}$.

Assigning Probabilities to Outcomes

- **Exercise 2:** balls numbered 1-100 are placed in a box. Four balls are chosen at random, 1 by 1 and *with replacement*. What is the probability of the outcome (53, 67, 12, 99)?
- An elementary event or outcome is an *ordered* sequence of 4 numbers.
- We need to count the number of possible ordered sequences of 4 numbers that can be drawn from a set of 100 numbers *with replacement*.
- **Counting formula 2:** The number of different ordered sequences of size k that can be drawn from n objects with replacement is

$$n^k.$$

- Therefore, $P(A = (53, 67, 12, 99)) = \frac{1}{100 \times 100 \times 100 \times 100} = \frac{1}{100,000,000}.$

Assigning Probabilities to Outcomes

- **Exercise 3:** balls numbered 1-100 are placed in a box. Four balls are chosen *simultaneously*. What is the probability of outcome $(53, 67, 12, 99)$?
- An elementary event or outcome is an *unordered* sequence of 4 numbers.
- We need to count the number of possible unordered sequences of 4 numbers that can be drawn from a set of 100 numbers.
- **Counting formula 3:** The number of different unordered sequences of size k that can be drawn from n different objects is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

- Therefore, $P(A = (53, 67, 12, 99)) = \frac{96!4!}{100!} = \frac{1}{3,921,225}.$

Assigning Probabilities to Outcomes

- **Example. Application of counting formulas.** Given that a family has 4 children and all possible sequences of boy and girl births are equally likely, indicate the probability of the sequence

bbbb.

- One possible way to answer this question is to list all possible sequences:
bbbb, bbbg, bbgb, bbgg, . . .
- A quicker way is to use the counting formulas we just learned.
- Note that the sequence *bgbg* is different from the sequence *bbgg*, so the order matters.
- Also note that we can think of a random phenomenon generating this sequence as taking a sequence of four balls, 1 by 1, with replacement, from an urn that contains only two balls: one ball with a “b” and one ball with a “g”.
- Using **counting formula 2**, we know that the total number of possible sequences is: $n^k = 2^4 = 2 \times 2 \times 2 \times 2 = 16$.

Assigning Probabilities to Outcomes

- Remember that counting formulas are useful to assign probabilities to **outcomes** in random phenomena in which **all outcomes in the sample space are equally likely**.
- If the outcomes of a random phenomenon are not equally likely, the only way to assign probabilities is to observe many repetitions of the random phenomenon and record the proportion of times that each outcome is observed.
- Independently of the way we assign probabilities to outcomes, these should respect certain rules:
 - The probability of an event is always positive or 0: $P(A) \geq 0$.
 - The probability of an event is always smaller or equal to 1: $P(A) \leq 1$.
 - The probability of the sample space is 1: $P(S) = 1$.
 - The probability of an event and its complement is 1: $P(A) + P(A^c) = 1$.
- These rules are consequences of the definition of probability as a proportion.

ASSIGNING PROBABILITIES TO EVENTS

Assigning Probabilities to Events

- As indicated above, an event is a set of outcomes.
- Once we know how to assign probabilities to outcomes, assigning probabilities to events is simple.
- **The probability of any event is the sum of the probabilities of the outcomes making up the event.**
- This is true independently of whether the outcomes are equally likely or not.
- Think of the random phenomenon “tossing a fair coin twice”. What is the probability of the event “obtaining at least one head”?
- The set of outcomes included in this event is

$$\{HH, HT, TH\}.$$

Therefore, its probability is

$$P(HH) + P(HT) + P(TH).$$

Assigning Probabilities to Events

- In the example in the previous slide, it was very easy to identify which of the outcomes of the random phenomenon “flipping a fair coin twice” are included in the event “obtaining at least one head”.
- The reason is that the sample space contains only 4 outcomes and we can go one by one testing whether it satisfies the condition that it contains at least one head.
- In random phenomena in which the sample space contains a very large number of outcomes, it can be quite hard to identify all the outcomes that are satisfied by one given event.
- **Example of counting outcomes included in an event.** In the baseball World Series, two teams play games until one team has won four games, thus the total length of the series must be between 4 and 7 games. What is the probability of each of these lengths under the assumption that the games are independent events with each team equally likely to win?

Assigning Probabilities to Events

- Using the counting rules, we can count the number of outcomes in the sample space. We can think of the World Series as a random phenomenon that consists in extracting 7 balls from an urn that contains only two balls (Y and R), with replacement.
- Note that the order matters. If the series is

$RRRRYYYY,$

it would have finished in 4 games. If the series is

$RYRYRYR,$

then it would have finished in 7 games.

- Therefore, the number of possible series is:

$$2^7 = 128.$$

Assigning Probabilities to Events

- How many of the 128 possible series imply a total length of the series of 4 games?
- Note that

RRRRXXX

implies that, independently of the value of X , the series finishes in 4 games and the Red Sox win. There are 2^3 series this kind.

- Analogously, there are also 2^3 of series in which the Yankees win in 4 games.
- Therefore, there are 2×2^3 possible outcomes such that the World Series finishes in 4 games.
- Therefore, the probability of the event “World Series finishes in 4 games” is

$$\frac{2^4}{2^7} = 2^{-3} = \frac{1}{8} = 0.125.$$

- The actual probability is 0.195. Why is our probability model wrong?

Assigning Probabilities to Events

- **Example.** Imagine that you are going for dinner with four of your classmates. The five of you have to decide between going to Despaña or going to Teresa. What is the probability that your vote will be decisive?
- For simplicity, let's assume that the vote of each of your four classmates is equally likely to go in each direction. Therefore, all the following events are equally likely:

$DDDD$	$DTDD$	$TDDD$	$TTDD$
$DDDT$	$DTDT$	$TDDT$	$TTDT$
$DDTD$	$DTTD$	$TDTD$	$TTTD$
$DDTT$	$DTTT$	$TDTT$	$TTTT$

- You will be a pivotal voter as long as 2 of your classmates vote for Despaña and 2 vote for Teresa. Therefore, the probability that you are pivotal is $6/16$.

Assigning Probabilities to Events

- **Example.** Now imagine that in your group of five going to dinner there are three women and two men.
- Consider the following voting rule: the voters of each sex will get together and form a pact; they will make a preliminary vote within their group and then, in the general vote of all 5 of you, they will all vote as a block in favor of the choice preferred by a majority of their sex.
- What is the probability that your vote is decisive if you are a man? What is the probability if you are a woman?
- Consider redoing the examples both in this slide and in the previous one if the probability that each individual votes for Despaña over Teresa is 0.6.

CONDITIONAL PROBABILITIES AND INDEPENDENCE

Conditional Probabilities and Independence

- Two events A and B are **independent** if knowing that one occurs does not change the probability that the other one occurs.
- **Example.** Think of the random phenomenon “throw a fair die with 20 sides”, marked $1, 2, \dots, 20$. Think of two events:

A = obtain an even number,

B = obtain a number equal to 10 or smaller.

- $P(A) = 1/2$. We compute this probability by listing the outcomes that make for event A , and dividing by the total number of outcomes.
- Now assume that we have obtained a number equal to 10 or smaller. What is the probability that this number is even?

$$P(A|B) = \frac{\{2, 4, 6, 8, 10\}}{\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}} = \frac{1}{2} = P(A)$$

- $P(A|B)$ is the conditional probability of A given B .

Conditional Probabilities and Independence

- Similarly,

$$P(B) = \frac{\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}}{\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20\}} = \frac{1}{2}$$

and

$$P(B|A) = \frac{\{2, 4, 6, 8, 10\}}{\{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}} = \frac{1}{2},$$

where $P(B|A)$ is the conditional probability of B given A .

- More formally, two events are **independent if and only if**

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B)$$

- Conditional probabilities are very useful to compute the probability that two events, A and B , happen together.

PROBABILITY OF A AND B

Multiplication Rule

- The probability that two events A and B happen together can be found as

$$P(A \text{ and } B) = P(A \cap B) = P(A|B)P(B) = P(B|A)P(A).$$

- Therefore, we can write the conditional probability as

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}.$$

- In the specific case in which A and B are independent,

$$P(A \text{ and } B) = P(A)P(B).$$

Multiplication Rule

- For more than two cases:

$$P(A \text{ and } B \text{ and } C) = P(A|B \text{ and } C)P(B|C)P(C)$$

- Example (same as in slide 17):** balls numbered 1-100 are placed in a box. Four balls are chosen at random, 1 by 1 and *without replacement*. What is the probability of the outcome (53, 67, 12, 99)?

$$\begin{aligned} P(\{53\} \cap \{67\} \cap \{12\} \cap \{99\}) &= \\ &= P(\{99\}|\{53\} \cap \{67\} \cap \{12\})P(\{53\} \cap \{67\} \cap \{12\}) \\ &= P(\{99\}|\{53\} \cap \{67\} \cap \{12\})P(\{53\}|\{67\} \cap \{12\})P(\{67\} \cap \{12\}) \\ &= P(\{99\}|\{53\} \cap \{67\} \cap \{12\})P(\{53\}|\{67\} \cap \{12\})P(\{67\}|\{12\})P(\{12\}) \\ &= \frac{1}{97} \times \frac{1}{98} \times \frac{1}{99} \times \frac{1}{100} \end{aligned}$$

- Note that we obtain the same probability that we have previously computed using the counting formula (see slide 17).

Conditional Probabilities and Independence

- **Example.** There are 3 cards in a hat. One of these cards is blue on both sides, one is pink on both sides, and one is blue on one side and pink on the other. Imagine that we select one card randomly and, without looking, we tape it to the blackboard. If the side facing the class is blue (pink), what is the probability that the other side of the card is also blue (pink)?
- Note that we are asking for a conditional probability:

$$P(\text{blue}_h | \text{blue}_v) = \frac{P(\text{blue}_h \text{ and } \text{blue}_v)}{P(\text{blue}_v)}$$

- Think that we have chosen the extracted card randomly and, given the random choice of the card, we have also chosen the visible side randomly.

Conditional Probabilities and Independence

- The probability that the visible side of the card is blue is 0.5 (i.e. there are the same number of blue and pink sides). Therefore

$$P(blue_v) = \frac{1}{2}$$

- The probability that both the visible and the hidden side of the card are blue is equal to the probability of drawing the card that is blue on both sides. Given that we select the card randomly, this probability is $1/3$. Therefore

$$P(blue_h \text{ and } blue_v) = \frac{1}{3}$$

- Therefore

$$P(blue_h | blue_v) = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Multiplication Rule

- **Example:** let's revisit the example in slide 20. In that case, because at that point we only knew how to compute the probability of events in random phenomena whose outcomes are all equally likely, we assumed that the probability of a boy at birth was the same as the probability of a girl at birth.
- In fact, that assumption is not right. In the US, every year from 1983 to 2013, there are approximately 1,050 boys born for every 1,000 girls. This implies that the probability of a boy at birth is

$$P(b) = \frac{1050}{1050 + 1000} \approx 0.51.$$

- Assuming that genders at birth are independent

$$P(bbbb) \approx 0.51^4$$

$$P(bgbg) \approx (0.51^2)(0.49^2)$$

- What is the probability of $P(bbgg)$?

PROBABILITY OF A OR B

Addition rule

- The addition rule allows to compute the probability that at least one of two events will happen: either the first happens, or the second, or both.

$$\begin{aligned}P(A \text{ or } B) &= P(A \cup B) = P(A) + P(B) - P(A \cap B) \\&= P(A) + P(B) - P(A|B)P(B)\end{aligned}$$

- **Example.** Let's revisit the random phenomenon in slide 30 but asking a different question. If you roll a die with 20 sides numbered 1 to 20, what is the probability that we either obtain an even number or a number smaller or equal to 10?

$$P(A \cup B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$$

BAYES RULE

Bayes Rule

- Bayes Rule allows to reverse the role of A and B in a conditional probability. That is, suppose $P(A|B)$ was known, how could you calculate $P(B|A)$?
- The formula is:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

and the proof is very simple:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}.$$

- Besides, using the multiplication rule,

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

and, therefore,

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

Bayes Rule

- **Example.** There are three prisoners. Two of the three are about to be released but only the warden knows which two. One prisoner asks the warden to secretly tell him the name of one of the other two prisoners who is to be released. He argues that since he knows that at least one of the others is to be released it won't change anything if he is given this information. But once the warden complies, the prisoner regrets his request, for he now thinks that his chances of being released have dropped from $2/3$ to $1/2$.
- Who thinks that the new probability is $1/2$?
- Who thinks that the new probability is $2/3$?
- (solution to be given on Thursday :-))

Bayes Rule

- **Example.** Is it possible that Lance Armstrong was innocent?
- Imagine a drug test that has 95% of accuracy.
 - An athlete that has consumed drugs will be tested positive in 95% of the cases.
 - An athlete that has not consumed drugs will be tested negative in 95% of the cases.
- Imagine that only 5% of the athletes really consume drugs.
- Given that Lance Armstrong was tested positive, what is the probability that he actually consumed drugs?

$$\begin{aligned} P(d|p) &= \frac{P(d \text{ and } p)}{P(p)} = \frac{P(p|d)P(d)}{P(p|d)P(d) + P(p|not\ d)P(not\ d)} \\ &= \frac{0.95 \times 0.05}{0.95 \times 0.05 + 0.05 \times 0.95} = \frac{1}{2} \end{aligned}$$

Bibliography

- MMC: *Introduction to the Practice of Statistics*
 - Chapters 4.1, 4.2, and 4.5.

WWS 507c: Quantitative Analysis

Lecture 3: Random Variables and Probability Distributions

Princeton University

September 18, 2014

RANDOM VARIABLES

Random Variables

- In Lecture 2, we introduced the concepts of
 - sample space: the set of possible outcomes of a random phenomenon;
 - outcome: each element of the sample space;
 - event: a set of outcomes or a subset of the sample space.

- E.g. the sample space of

- rolling a die is

$$S_1 = \{\text{obtain one dot at the top, obtain two dots at the top, } \dots\};$$

- simultaneously flipping two coins, one dollar and one euro, is

$$S_2 = \{HH, HT, TH, TT\},$$

where HT implies a head on the dollar and a tail on the euro.

- simultaneously flipping two identical coins is

$$S_3 = \{HH, HT, TT\},$$

where HT denotes that one of them is heads and the other one is tails.

Random Variables

- Outcomes are not numbers.
- Think about the random phenomenon “obtaining a random sample of 5 students from the population of students taking WWS507c”.
- One possible outcome from this phenomenon is:

$\{Ashley, Ann, Connor, Jack, Brendan\}$.

- Another possible outcome would be

$\{John, Dustin, Kabira, Daniel, Tommaso\}$.

- However, in most cases, I would not be interested in the names or identities of the people in my random sample. What I might care about is things like:
 - how many of them like my class?
 - how many of them read at home the slides from Lecture 2 (...ehem, ehem...)?
- Note that the things I care about are **numbers**.

Random Variables

- Random variables are translations of outcomes of a random phenomenon into numbers. Random variables assign numbers to all the elements in the sample space of a random phenomenon.
- For example, if I were to be interested in how many of you guys have read the slides from Lecture 2, my random variable would translate

$\{Ashley, Ann, Connor, Jack, Brendan\}$.

into a number 5, and

$\{John, Dustin, Kabira, Megan, Tommaso\}$.

also into a number 5.

- Note that random variables might assign the same number to different outcomes of a random phenomenon!

Random Variables

- Random variables are usually denoted with letters like X, Y, and Z.
- Formally, random variables are functions that map the sample space of a random phenomenon into real numbers.
- We can define different random variables on the sample space.
- Again imagine that the random phenomenon is “obtaining a random sample of 5 students from the population of students taking WWS507c”.
- As we have discussed, the sample space for this random phenomenon is the set of all possible groups of 5 students that I can form from the larger group of all students taking WWS507c.
- On this sample space, we can define many variables
 - X: number of female students

$$X(\{Ashley, Ann, Connor, Jack, Brendan\}) = 2$$

- Y: number of names that start with an A

$$Y(\{Ashley, Ann, Connor, Jack, Brendan\}) = 2$$

Random Variables

- Just to be sure we understand what is going on, let's describe two more random variables but using a different random phenomenon.
- Think of the random phenomenon "flipping two distinguishable coins".
- One random variable we might define is
 - X : number of heads.

S	P(S)	X
TT	1/4	0
HT	1/4	1
TH	1/4	1
HH	1/4	2

or

$$\{X = 0\} = \{TT\}, \quad \{X = 1\} = \{HT, TH\}, \quad \{X = 2\} = \{HH\}.$$

The random variable X can take three values, 0, 1, and 2, which are events defined by specific collections of outcomes.

Random Variables

- Another variable that we can define on the same sample space is
 - Y : variable equal 1 if both coins have identical symbol.

S	P(S)	Y
TT	1/4	1
HT	1/4	0
TH	1/4	0
HH	1/4	1

or

$$\{Y = 0\} = \{HT, TH\}, \quad \{Y = 1\} = \{HH, TT\}.$$

The random variable Y can take two values, 0, and 1, which are events defined by specific collections of outcomes,.

Random Variables

- Finally, let's find another example that is more relevant for public policy.
- The official unemployment rate in the US is computed by the Bureau of Labor Statistics using the Current Population Survey, a monthly household survey conducted by the Bureau of the Census.
- Let's imagine that this survey is done by a sample of 60,000 people drawn from all US civilians 16 years and older.
- Let's imagine that the official unemployment rate is computed as the share of these 60,000 people that report being unemployed.
- What is the random phenomenon behind this bureaucratic process?
- What is the sample space?
 - How would you compute the total number of elements in the sample space?
- Define a random variable that will give you the official unemployment rate.
 - How many different values this random variable might take?

PROBABILITIES DEFINED OVER RANDOM VARIABLES: PROBABILITY DISTRIBUTION

Probability Distribution

- In Lecture 2, we defined the probability for each outcome and we constructed probabilities for events.
- Specifically, using the addition rule, we saw that the probability of an event A is equal to the sum of the probabilities of all outcomes included in event A .
- Example. In the random experiment flipping a coin twice, the probability of

$A = \text{obtaining at least one head}$

is equal to

$$P(A) = P(HT) + P(HH) + P(TH) = \frac{3}{4}$$

- But note that this is also the probability of observing $X = 1$ if we define X as a random variable that takes value 1 if we obtain at least one head:

$$P(X = 1) = \frac{3}{4}.$$

- The set of outcomes that make X equal 1 is just an event for the random phenomenon “flipping a coin twice”.

Probability Distribution

- More generally, once we have defined a random variable for a particular random phenomenon, we can actually compute the probability that this random variable takes any possible value.
- Let's go back to the example in which we defined a random variable

$$X = \text{number of heads}$$

over the random phenomenon "flipping two distinguishable coins". Then:

S	P(S)	X		X	P(X)
TT	1/4	0	→	0	1/4
HT	1/4	1		1	1/2
TH	1/4	1		2	1/4
HH	1/4	2			

X can take three values, 0, 1, and 2. The probability that X takes these values is $P(X = 0) = P(X = 2) = 0.25$, and $P(X = 1) = 0.5$.

Probability Distribution

- Given a random phenomenon, the set of all possible values that a random variable may take when applied to that random phenomenon is called the **support of the random variable**.
- The **probability distribution** of a random variable X lists all the values in its support and the probability that X takes each of these values.
- In the example in the previous slide, the probability distribution is

X	$P(X)$
0	$1/4$
1	$1/2$
2	$1/4$

Probability Distribution

- In Lecture 2, we said that what fully characterizes a random phenomenon is
 - its sample space,
 - the probability of each outcome in the sample space.
- If the only thing we care about in a random phenomenon is the information contained in a random variable X , then the only thing we want to know about such random phenomenon is the probability distribution of X :
 - the set of possible values that X might take (support of X),
 - the probability that X takes each value in its support.
- In a random phenomenon, if we only care about the information captured in X , the probability distribution of X extracts all the information that is relevant to us.
- The sample space and the probability of each outcome in the sample space is really there, lurking in the background, but they provide too much information. The probability distribution of X summarizes all the information about a random phenomenon that is relevant for X .

Probability Distribution

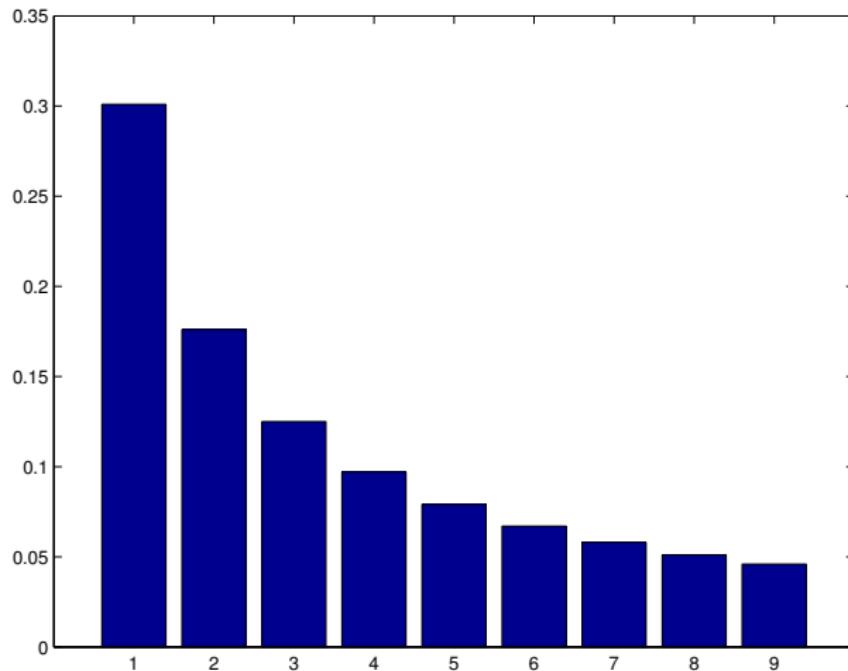
- Let's get some practice computing probability distribution functions.
- **Example.** A supervisor has two men and two women working for her. She wants to choose two workers for a special job. Not wishing to show any biases in her selection, she decides to select the two workers at random. Let X denote the number of women in her selection. Find the probability distribution for X .
- **Example.** Imagine that I throw two six-sided dice, one blue and one red. Compute the probability distribution of a variable X = “difference between the number in the red die and the number in the blue die”. Compute also the probability distribution of a variable Y that takes value 1 if the blue die gets a larger number than the red one.

Probability Distribution

- Guess which is the first digit of the address of my parents' house in Madrid.

Benford's Law

- The most famous probability distribution is Benford's Law

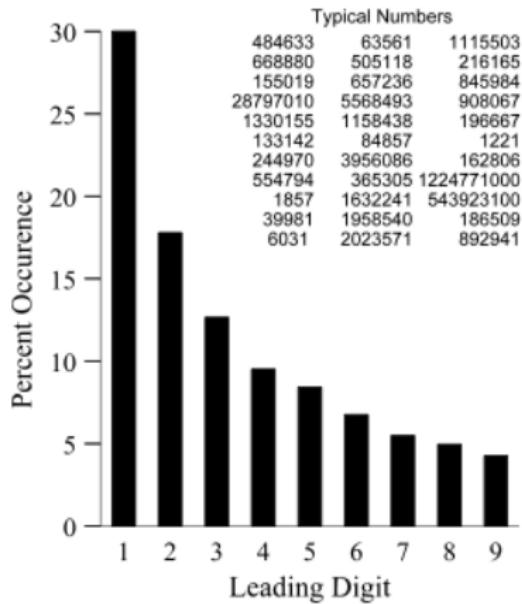


Benford's Law

- Dr. Frank Benford, a physicist at General Electric in the 1930s noticed that, in books of logarithms, the pages corresponding to numbers with a leading digit of 1 were more worn than other pages.
 - This was first noticed by Simon Newcomb in 1881, but Benford rediscovered it
- Benford inferred that numbers beginning with the digit 1 were being looked more often than numbers beginning with 2 through 9.
- Benford seized upon this idea and spent years collecting data to show that this pattern was widespread in nature. In 1938, he published his results, citing more than 20,000 values such as numbers in magazines, baseball statistics, areas of rivers, population of countries, etc. The resulting probability distribution for first digits came to be known as Benford's law.
- If you do not believe it, go through several pages of today's newspaper and examine the first or leading digit of each number. Count the frequency of times that you find each number between 1 and 9. You will find that about 30.1% of numbers will start with 1!

Benford's Law

- Even more fun...if you compute the probability distribution of the leading digit for a sufficiently large set of numbers randomly taken from the U.S. Federal income tax returns, you will obtain



Benford's Law

- If you convert all these dollar values to euros using today's conversion rate, it is likely that the leading digit of all these US Federal income tax returns will be changed by this conversion.
- Nevertheless, about 30.1% of the converted numbers will still have a leading digit of 1!
- In other words, if a set of numbers follows Benford's law, multiplying the numbers by any possible constant will create another set of numbers that also follows Benford's law.
- Benford's Law has been used in forensic accounting and auditing as an indicator of accounting and expenses fraud. For more info on this, there is a great *Radio-Lab* episode:
<http://www.radiolab.org/story/91699-from-benford-to-erdos/>
- However, apparently you should not use it to detect fraud in elections:
http://vote.caltech.edu/sites/default/files/benford_pdf_4b97cc5b5b.pdf

COMMON PROBABILITY DISTRIBUTIONS

Common Probability Distributions

- Benford's Law is an amazing empirical regularity.
- However, its applicability is limited because we do not know yet which are the characteristics that random phenomena have to verify so that some random variable follows the probability distribution described by Benford's Law.
- Contrary to Benford's Law, there are other equally famous probability distributions for which we understand the characteristics that random phenomena must have so that some random variable may be described according to these distributions.
- Some of these distributions are:
 - binomial distribution,
 - geometric distribution,
 - negative binomial distribution.

Binomial Distribution

- In Lecture 2, we saw that Ray Allen has a 40% probability of scoring a three-pointer. Which is the probability that he scores 7 three-pointers if he tries 7 times? Assume that each shot is independent of the previous ones.
- In US, every year from 1983 to 2013, there are approximately 1,050 boys born for every 1,000 girls. This implies that the probability of a boy at birth is

$$P(b) \approx \frac{1050}{1050 + 1000} \approx 0.51$$

Assume that genders at birth are independent. What is the probability that a family has three girls and one boy?

- And what is the probability that in a WWS507c class with 43 students there are 20 women?
- If I read to you a sequence of 7 random digits, the probability that, after 15 seconds, you remember the sequence is approx. 0.9. If I were to read such sequence to all of you, what is the probability that all of you remember the sequence correctly?

Binomial Distribution

- All these random variables have something in common. They all count the number of “successes” during a random phenomenon that has the following characteristics:
 - it consists of a fixed number of trials,
 - the outcome of each trial may be “success” or “fail”,
 - the probability of success in each of the trials is constant across trials,
 - the trials are independent.
- Binomial variables do not take into account the ordering of the outcomes.
- Do you think that all the four examples in the previous slide verify these four conditions?
- Denoting the number of trials by n and the probability of each trial by p , the probability distribution of a binomial random variable Y is:

$$p(Y = y) = \frac{n!}{(n - y)!y!} p^y (1 - p)^{n-y}$$

where y is a particular number between 0 and n .

Binomial Distribution

- **Example.** A lot of 5000 cars contains 5% defectives. If we select a 100 cars randomly, what is the probability that we observe no defective car?

$$P(Y = 0) = \frac{100!}{100!0!} 0.05^0 0.95^{100} = 0.95^{100} = 0.0059$$

What is the probability that we observe 10 defective cars?

$$P(Y = 10) = \frac{100!}{90!10!} 0.05^{10} 0.95^{90} = 0.0167$$

Geometric Distribution

- In Spain, people are usually late.
- People are so late that, if you are meeting with a friend at 8pm, the probability that this friend shows up in any minute after 8pm is constant at 1%. Just to be clear, this means that the probability that your friend shows up between 8 and 8:01pm is 1%, the probability that he shows up between 8:01 and 8:02pm (given that it has not shown up before 8:01pm) is 1%, the probability that he shows up between 8:02 and 8:03pm (given that it has not shown up before 8:02pm) is 1%, and so on.
- What is the probability that your friend shows up between 8 and 8:01pm?
- What is the probability that your friend shows up between 8:15pm and 8:16pm?
- What is the probability that your friend shows up between 8:30 and 8:31pm?

Geometric Distribution

- A geometric random variable Y is defined as the number of independent and identical success/failure trials until the first success occurs.
- Similarity with the binomial distribution: the geometric distribution is also related to a random phenomenon that consists in a sequence of independent success/failure trials in which the probability of success is constant from trial to trial.
- Difference with the binomial distribution:
 - binomial random variable: number of success in a fixed number n of trials.
 - geometric random variable: number of trials until the first success occurs
- Denoting the probability of success in each trial by p , the probability distribution of a geometric random variable is

$$p(y) = P(Y = y) = (1 - p)^{y-1} p,$$

where y is any number larger or equal to 1.

Geometric Distribution

- The probability that your friend shows up between 8 and 8:01pm is

$$P(Y = 1) = 0.01.$$

- The probability that your friend shows up between 8:15 and 8:16pm is

$$P(Y = 16) = 0.99^{15}0.01 = 0.0086.$$

- The probability that your friend shows up between 8:30 and 8:31pm is

$$P(Y = 31) = 0.99^{30}0.01 = 0.0074.$$

- As you can see, given that your friend has not showed up yet, the probability that he shows up in the next minute is independent of how much you have been waiting so far. However, the probability that you end up waiting y minutes decreases as y increases.

Geometric Distribution

- My dad is always on time. Sadly, my mum is not. So my dad has started to implement the following policy: if my mum has not showed up before five minutes after the time at which they were meeting have gone by, then my dad goes to the closest bar and waits there having a beer and watching soccer. If the probability that my mum shows up in any minute after the scheduled time is constant at 1%, what is the probability that my mum meets my dad at the bar?
- We know that my dad will be at the bar unless my mum shows up during the first four minutes after the scheduled time. Therefore

$$\begin{aligned}P(\text{bar}) &= 1 - P(Y = 1) - P(Y = 2) - P(Y = 3) - P(Y = 4) \\&= 1 - 0.01 - 0.99 \times 0.01 - 0.99^2 \times 0.01 - 0.99^3 \times 0.01 \\&= 1 - 0.01 \times (1 + 0.99 + 0.99^2 + 0.99^3) = 0.9606\end{aligned}$$

Negative binomial random variable

- A negative binomial random variable Y is defined as the number of independent and identical success/failure trials until the r th success occurs.
- The geometric random variable is a special case of the negative binomial random variable. In the former, $r = 1$.
- The negative binomial probability distribution is

$$p(Y = y) = \frac{y-1}{(y-r)!(r-1)!} p^r (1-p)^{y-r}, \quad y = r, r+1, r+2, \dots,$$
$$p(Y = y) = 0, \quad y < r.$$

CUMULATIVE DISTRIBUTION FUNCTION

Cumulative distribution function of a discrete random var.

- While the probability distribution function indicates the probability that a random variable Y takes any individual possible value y , the cumulative distribution function (CDF) shows the probability that a random variable Y is less *or equal* to a particular value y .
- While we usually use $p(y)$ to denote the probability distribution function, we usually denote the CDF as $F(y)$:

$$F(y) = P(Y \leq y) = \sum_{y_i \leq y} P(Y = y_i)$$

- **Exercise.** Draw the probability distribution and cumulative distribution functions for the random variable “number of observed dots when throwing a fair dice once”.
- Note that
 - $0 \leq F(y) \leq 1$.
 - $F(y)$ is nondecreasing on y : $F(y_1) \leq F(y_2)$ when $y_1 \leq y_2$.
 - For any y_1 and y_2 : $P(y_1 < Y \leq y_2) = F(y_2) - F(y_1) = \sum_{y_1 < y \leq y_2} P(Y = y)$.

Bibliography

- MMC: *Introduction to the Practice of Statistics*
 - Chapters 4.3.

WWS 507c: Quantitative Analysis

Lecture 4: Random Variables and Probability Distributions (cont.)

Princeton University

September 23, 2013

REVIEW OF LECTURE 2

Review of Lecture 2

- Think of the random phenomenon “roll a blue and red die”.
 - Write the sample space.
 - Write the probability of each outcome in the sample space.
 - Write the support of a random variable X defined as

Outcome of blue dice minus outcome of red dice

- Write the probability distribution function for X .
- Write the cumulative distribution function for X .

CONTINUOUS RANDOM VARIABLES

Continuous vs. Discrete Random Variables

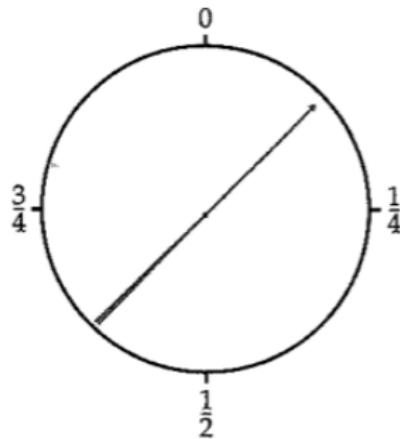
- In Lecture 2, we focused on variables that can take only a finite or countably infinite number of distinct values.
 - more formally, variables whose support is a finite or countably infinite set
- Examples
 - first or leading digit of a number;
 - number of unemployed workers in the U.S.;
 - number of claps that are heard in a theater after a show;
 - dummy value for being left-handed;
- We denote these variables as **discrete random variables**.
- The number of possible values that these random variables take may be small (like in example 1) or very large (like in example 2). However, in every case, it is conceptually possible to list all the possible values that these random variables might take (we can list all the values in their support)

Continuous vs. Discrete Random Variables

- There are many random variables that do not fall into the category of discrete random variables.
- Think of a random variable measuring the average daily rainfall in Princeton during the year 2013.
 - Theoretically, with measuring equipment of perfect accuracy, the amount of rainfall could take on any value between 0 and 100 inches.
- Think of a random variable measuring the time you have to wait until the NJ transit arrives.
 - Theoretically, this time measured in minutes can be any number between 0 minutes and (sadly) ∞ .
- Note that one can measure $3.456785\dots$ inches of rain (as long as the equipment is precise enough) and one can measure a time lag of $25.7689\dots$ minutes until the NJ transit arrives (as long as one has an atomic watch).

Continuous Random Variables

- Think of a spinner that turns on its axis. The pointer can come to rest anywhere on a circle that is marked from 0 to 1.



- The sample space is now an entire interval of numbers:

$$S = \{\text{all numbers } x \text{ such that } 0 \leq x \leq 1\}$$

PROBABILITY DENSITY FUNCTION

Probability Density Function

- The prob. that a continuous random variable takes any particular value is 0.
- Therefore, asking questions like
 - what is the prob. that it takes exactly 56 mins. to go to the airport?, or
 - what is the prob. that the max. temperature today is exactly 67 degrees?

is meaningless.

- For continuous random variables, only intervals of values have positive probability.
 - What is the probability that the spinner stops at a point higher than 0.3 and lower than 0.7?

$$Pr\{0.3 \leq x \leq 0.7\}$$

- From the definition of probability, remember that what we are really asking is: *if we were to turn the spinner many many times, in what fraction of them would the spinner stop at a point between 0.3 and 0.7?*

Probability Density Function

- For discrete variables, we saw in Lecture 2 that the **probability distribution** of a random variable X is a mathematical object that allows us to compute the probability that X takes **any** subset of values in its support.
- As an example, using the probability distribution you computed in slide 2

$S(X)$	-5	-4	-3	-2	-1	0	1	2	3	4	5
$P(X)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

- Using the information contained in this **probability distribution** we can compute probabilities like
 - $P(X = 4)$
 - $P(X = 2.5)$
 - $P(-2 \leq X \leq 2)$
 - $P(\{X \leq -3\} \cup \{X \geq 2\})$
 - and so on...

Probability Density Function

- For continuous random variables, we would also like to have a mathematical object that allowed us to easily compute the probability that a random variable X takes values in **any** given interval x_1 to x_2

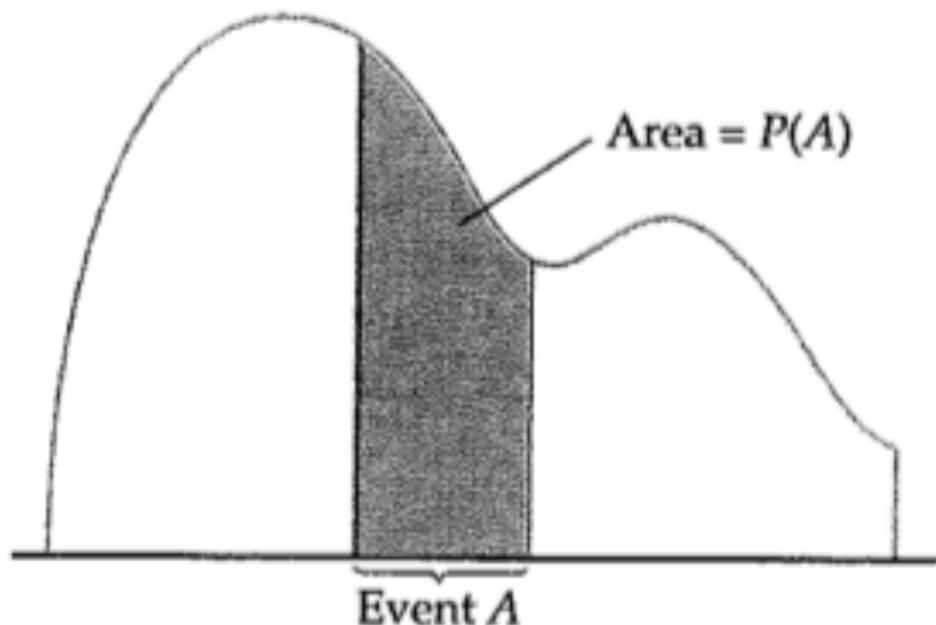
$$Pr(x_1 \leq X \leq x_2).$$

- This mathematical object exists! It is the **probability density function** (PDF).
- For each random variable X , the probability density function of X is a function $f(X)$ such that, for any two values x_1 and x_2 , the probability

$$Pr(x_1 \leq X \leq x_2)$$

is equal to the area of the space limited by two vertical lines at x_1 and x_2 , a horizontal axis at 0, and the function $f(X)$.

Probability Density Function



Small Math Review

- Note that using the probability density function to compute probabilities just converts a problem of computing a probability like

$$Pr(x_1 \leq X \leq x_2)$$

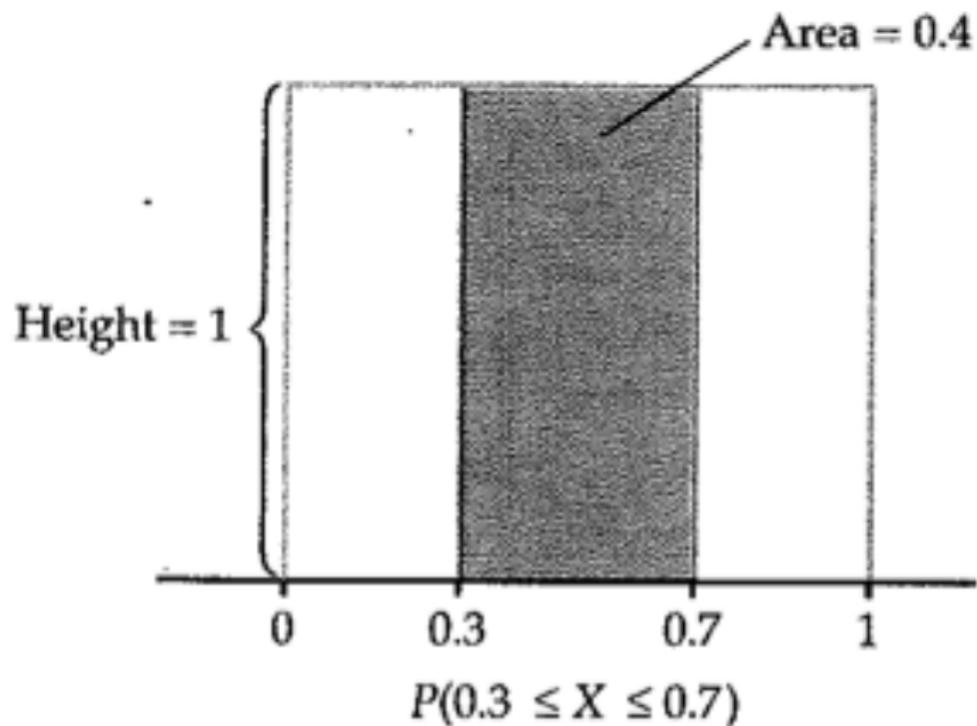
into a problem of computing an area under a function.

- How do we compute areas under functions? As an example, how do we compute the area of the shaded figure in slide 12?
- We use integrals!
- The integral of a function $f(X)$ with lower limit x_1 and upper limit x_2 is equal to the area under $f(X)$ and limited by a horizontal line at $f(X) = 0$, and two vertical lines at $X = x_1$ and $X = x_2$.
- Therefore,

$$Pr(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(X)dx$$

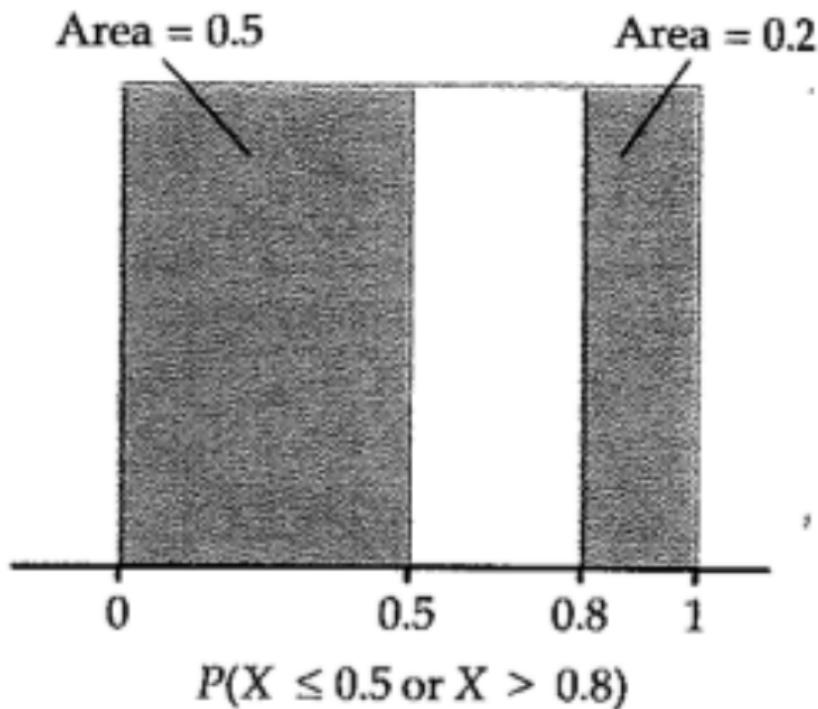
Computing Probabilities

- Example 1.



Computing Probabilities

- Example 2.



Computing Probabilities

- Note that all the information about a random variable X is contained in its probability density function!
- Once you know the probability density function of X , you can compute the probability of any event that refers to X .
- **Example** Assume that the density function of some random variable X is

$$f(x) = \begin{cases} 0, & x \leq 0, \\ x, & 0 > x \leq 1, \\ 2 - x, & 1 < x \leq 2, \\ 0, & x > 2. \end{cases}$$

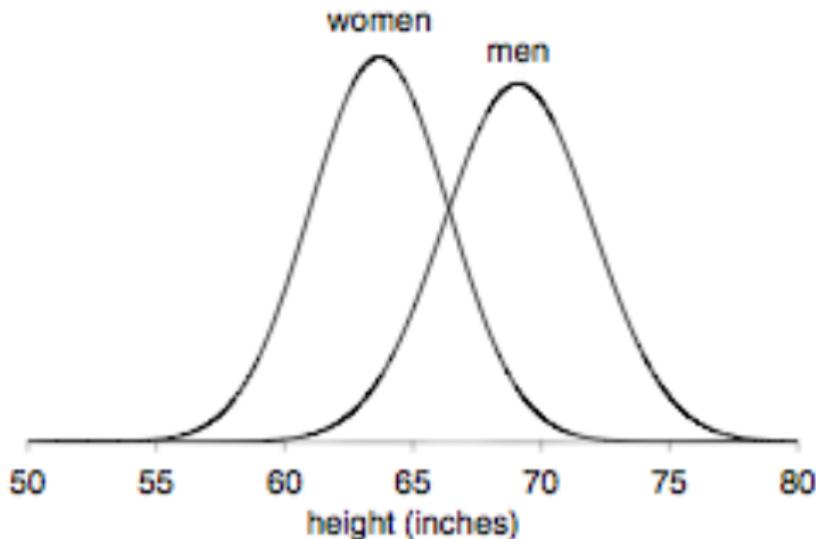
Compute:

- $P(-0.5 \leq X \leq 0.5)$
- $P(X \leq 1)$
- $P(X > -2)$

COMMON PROBABILITY DENSITY FUNCTIONS

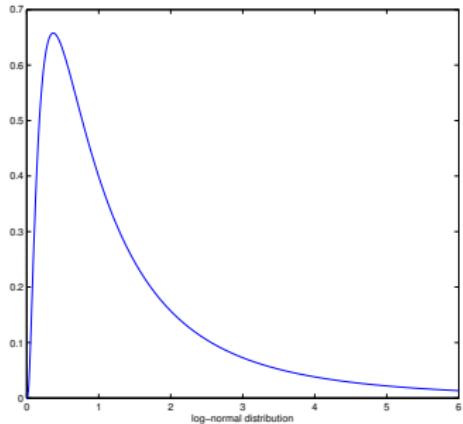
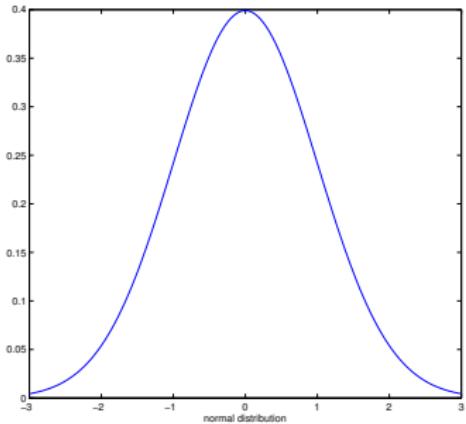
Normal Random Variable

- One example of a continuous random variable is the normal random variable.
- As an example, the distribution of heights of adult men and women in the United States follows a distribution that approximately normal.



Log-normal Random Variable

- If X is normally distributed, then $Y = \exp(X)$ is log-normally distributed.
- Examples of variables that follow close to log-normal distributions:
 - household income within a country;
 - age of first-marriage in Western countries;
 - length of spoken words in phone conversation in the US;
 - number of words per sentence written by G. K. Chesterton and G. B. Shaw;



CUMULATIVE DISTRIBUTION FUNCTION

Cumulative Distribution Function

- The cumulative distribution function (CDF) of a continuous random variable is defined in exactly the same way as for a discrete random variable.
- The value of the CDF at a particular point in the support of a random variable X (let's call this point x) is equal to the probability that the random variable X takes any value smaller than x

$$F(b) = P(X \leq b).$$

- Rewriting $P(X \leq x)$ as a function of the probability density function of X , we can express $F(b)$ as

$$F(b) = \lim_{a \rightarrow -\infty} \int_a^b f(X) dX$$

- In order to go from the PDF to the CDF, we need to integrate!
- In order to go from the CDF to the PDF, we need to differentiate!

$$f(a) = \frac{\partial F(X)}{\partial X} \Big|_{x=a}$$

- **Example.** A random variable whose PDF is flat between two numbers a and b is called **uniformly distributed**. An example of a uniformly distributed random variable is:

$$f(x) = \begin{cases} 0, & x \leq -1, \\ \frac{1}{2}, & -1 > x \leq 1, \\ 0, & x > 1. \end{cases}$$

The CDF of this random variable X is

$$F(x) = \begin{cases} 0, & x \leq -1, \\ \frac{x}{2} + \frac{1}{2}, & -1 > x \leq 1, \\ 1, & x > 1. \end{cases}$$

Note that, if $-1 > a \leq 1$, then

$$F(a) = \int_{-\infty}^a f(X)dX = \int_{-\infty}^{-1} 0dX + \int_{-1}^a \frac{1}{2}dX = 0 + \left[\frac{X}{2} \right]_1^a = \frac{a}{2} - \frac{-1}{2} = \frac{a}{2} + \frac{1}{2}$$

PDF and CDF

- The PDF and the CDF of a random variable X contain exactly the same information about X .
- For any two real numbers a and b , we can compute

$$P(a \leq X \leq b)$$

either using the probability density function of X

$$P(a \leq X \leq b) = \int_a^b f(X) dX$$

or using the cumulative density function of X

$$\begin{aligned} P(a \leq X \leq b) &= \int_a^b f(X) dX = \int_{-\infty}^b f(X) dX - \int_{-\infty}^a f(X) dX \\ &= F(b) - F(a) \end{aligned}$$

Properties of the PDF and CDF

- Given that, for any real numbers a and b ,

$$P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(X)dX,$$

and

$$0 \leq P(a \leq X \leq b) \leq 1,$$

then,

- 1 $f(x) \geq 0$ for every x ,
- 2 $F(x) \geq 0$ for every x ,
- 3 $F(x)$ is always weakly increasing,
- 4 $\lim_{x \rightarrow -\infty} F(x) = F(-\infty) = 0$,
- 5 $\lim_{x \rightarrow \infty} F(x) = F(\infty) = 1$,
- 6 $F(b) - F(a)$ is the area under the PDF between a and b .
- 7 The area under the PDF between $-\infty$ and ∞ is 1.

Example

- Assume, for some constant A ,

$$f(y) = \begin{cases} 0, & y < 0, \\ y, & 0 \leq y \leq A, \\ 0, & y > A, \end{cases}$$

- Compute the CDF at some value y such that $0 \leq y \leq A$,

$$F(y) = \int_{-\infty}^y f(t)dt = \int_0^y f(t)dt = \int_0^y tdt = \frac{t^2}{2} \Big|_0^y = \frac{y^2}{2}$$

- Compute the value of A that is consistent with $f(y)$ being the PDF of some random variable

$$F(A) = 1, \quad \rightarrow \quad A = \sqrt{2} \quad \left(\frac{\sqrt{2}^2}{2} = 1 \right)$$

- Compute $P(0.5 \leq Y \leq \sqrt{2})$

$$P(0.5 \leq Y \leq \sqrt{2}) = F(\sqrt{2}) - F(0.5) = 1 - \frac{0.5^2}{2} = 1 - \frac{0.25}{2} = 0.875.$$

Bibliography

- MMC: *Introduction to the Practice of Statistics*
 - Chapters 4.3.

WWS 507c: Quantitative Analysis

Lecture 5: Random Variables and Probability Distributions (cont.)

Princeton University

September 25, 2014

REVIEW OF LECTURES 3 AND 4

Review of Lectures 3 and 4

- In lectures 3 and 4,
 - we introduced the concept of random variable;
 - we learned to differentiate between discrete and continuous random variables;
 - we saw how to use the probability distribution function (PDF) and the cumulative distribution function (CDF) to compute the probability that a discrete random variable Y takes some given set of values y_1, y_2, \dots .

$$Pr(\{Y = y_1\} \cup \{Y = y_2\}) = P(Y = y_1) + P(Y = y_2)$$

- we saw how to use the probability density function (PDF) and the cumulative distribution function (CDF) to compute the probability that a continuous random variable Y takes some value in a given interval $[y_1, y_2]$

$$Pr(y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} f(Y)dY = F(y_2) - F(y_1)$$

- In lectures 3 and 4, we treated each random variable in isolation. In this lecture, we will focus on the study of relationships between random variables.

DISCRETE RANDOM VARIABLES: JOINT PROBABILITY DISTRIBUTION

Joint Probability Distribution

- The Secretary of Defense has to decide on the total size of the US army for the incoming year. He is concerned with the possibility that a military conflict starts either in country A or in country B. If a conflict breaks in country A, he will need to send 10,000 soldiers to country A. If a conflict starts in country B, he will need to send 10,000 soldiers to country B. The Secretary of Defense would like to have the smallest possible army subject to the constraint that, no matter what happens, there is no shortage of soldiers.
- Define a random variable X_A that takes value 1 if there is a conflict in country A (and 0 if there is not). Analogously, define a random variable X_B that takes value 1 if there is a conflict in country B (and 0 if there is not).
- Assume that

$$P(X_A = 1) = 0.5$$

$$P(X_B = 1) = 0.5$$

- Is this enough information for the Secretary of Defense?

Joint Probability Distribution

- No. The Secretary of Defense cares about the probability that there is a conflict both in country A **and** in country B at the same time.
- If the probability that there is a conflict both in country A and in country B at the same time is zero, then the Secretary of Defense will plan to have an army of only 10,000 soldiers. If this probability is positive, then he will plan to have an army of 20,000 soldiers.
- In this setting, the PDF of X_A and the PDF of X_B do not provide enough information. This is because the PDFs of each of the two random variables do not provide any information about the degree of association of these two random variables. They do not allow to compute the probability that $X_A = 1$ **and** $X_B = 1$.

$$Pr(\{X_A = 1\} \cap \{X_B = 1\})$$

- The probability distribution that contains this extra information is called the **joint probability distribution** of X_A and X_B .

Joint Probability Distribution

- Given a pair of discrete random variables X_A and X_B , the **joint probability distribution** is the probability that X_A takes value x_A and X_B takes value x_B , for any possible value of (x_A, x_B) :

$$\begin{aligned} p(x_A, x_B) &= \Pr(\{X_A = x_A\} \cap \{X_B = x_B\}) = \Pr(X_A = x_A \text{ and } X_B = x_B) \\ &= \Pr(X_A = x_A, X_B = x_B). \end{aligned}$$

- In our example, one possible joint probability distribution of X_A and X_B is

		$X_B = 0$	$X_B = 1$
$X_A = 0$	0	0.5	
	$X_A = 1$	0.5	0

- In this case, the probability that both countries A and B suffer a military conflict in the same year is 0. The probability that none of them suffers a military conflict is also 0. Therefore, the army should have 10,000 soldiers and they will all be employed with probability one.

Joint Probability Distribution

- Another possible joint probability distribution of X_A and X_B is

		$X_B = 0$	$X_B = 1$
$X_A = 0$	0.5	0	
	0	0.5	

How many soldiers would be necessary in this case?

- Still another possibility would be

		$X_B = 0$	$X_B = 1$
$X_A = 0$	0.25	0.25	
	0.25	0.25	

- In all 3 joint probability distributions above, it is true that $P(X_A = 1) = 0.5$ and $P(X_B = 1) = 0.5$. The only feature that changes across them is the dependence between the random variables X_A and X_B .

Example

- Let's look at a second example that uses actual data.
- Define two variables describing characteristics of civilian non-institutional employed workers 25 years old and over.
- Variable X indicates level of education.

$$X = \begin{cases} 1, & \text{if less than a high school diploma,} \\ 2, & \text{if high school graduates, no college,} \\ 3, & \text{if less than a bachelor's degree,} \\ 4, & \text{if college graduates.} \end{cases}$$

- Variable Y indicates occupation.

$$Y = \begin{cases} 1, & \text{if managerial, professional and related,} \\ 2, & \text{if service,} \\ 3, & \text{if sales and office,} \\ 4, & \text{if natural resources, construction, and maintenance,} \\ 5, & \text{if production and transportation.} \end{cases}$$

Example

- Details on the definition of each occupation are contained in:
<http://www.bls.gov/cps/cenocc2010.pdf>
- Using data from the Current Population Survey (U.S. Bureau of Labor Statistics) for 2010, we can build the joint distribution function of X and Y .
- Guess the 16 numbers that characterize the joint distribution of X and Y :

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$
$X = 1$					
$X = 2$					
$X = 3$					
$X = 4$					

Example

- Expressed in percentages, the actual joint distribution of X and Y is:

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$
$X = 1$	0.5	2.7	1.1	1.9	2.1
$X = 2$	4.6	6.0	7.6	4.2	5.7
$X = 3$	9.1	4.8	8.1	2.7	3.0
$X = 4$	25.8	2.2	6.1	0.8	1.0

- The probability that a worker has less than a high school diploma and has a managerial occupation is 0.5%.
- The probability that a worker has a college degree and works in construction or maintenance is 0.8%.
- The probability that a worker has a college degree and a managerial occupation is 25.8%.

DISCRETE RANDOM VARIABLES: JOINT CUMULATIVE DISTRIBUTION FUNCTION

Joint Cumulative Distribution Function

- The **joint cumulative distribution function** (CDF) for a random vector (X, Y) evaluated at a point (x, y) indicates the probability that X is smaller or equal than x and Y is smaller or equal than y .

$$F(x, y) = P(\{X \leq x\} \cap \{Y \leq y\}) = P(X \leq x, Y \leq y).$$

- For discrete random variables, we can compute the joint CDF using the joint probability distribution function:

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_i \leq y} p(x_i, y_i).$$

Example

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$
$X = 1$	0.5	2.7	1.1	1.9	2.1
$X = 2$	4.6	6.0	7.6	4.2	5.7
$X = 3$	9.1	4.8	8.1	2.7	3.0
$X = 4$	25.8	2.2	6.1	0.8	1.0

- The joint CDF evaluated at $X = 2$ and $Y = 2$

$$F(X = 2, Y = 2) = F(2, 2)$$

indicates the probability that a worker has a high school degree or less and is employed in a managerial occupation or in services:

$$\begin{aligned}F(2, 2) &= p(1, 1) + p(1, 2) + p(2, 1) + p(2, 2) \\&= 0.5 + 4.6 + 2.7 + 6.0 = 13.8\%\end{aligned}$$

Example

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$
$X = 1$	0.5	2.7	1.1	1.9	2.1
$X = 2$	4.6	6.0	7.6	4.2	5.7
$X = 3$	9.1	4.8	8.1	2.7	3.0
$X = 4$	25.8	2.2	6.1	0.8	1.0

- The joint CDF evaluated at $X = 4$ and $Y = 1$

$$F(X = 4, Y = 1) = F(4, 1)$$

indicates the probability that a worker is employed in a managerial occupation (independently of her education)

$$\begin{aligned}F(4, 1) &= p(1, 1) + p(2, 1) + p(3, 1) + p(4, 1) \\&= 0.5 + 4.6 + 9.1 + 25.8 = 40\%\end{aligned}$$

DISCRETE RANDOM VARIABLES: MARGINAL PROBABILITY DISTRIBUTION

Marginal Probability Distribution

- The **marginal probability distribution** of a random variable X lists all the values in its support and the probability that X takes each of these values.
- The marginal probability distribution of a random variable X can be computed from the joint distribution of X with any other random variable Y :

$$p_X(X = x) = p_X(x) = \sum_{\text{all } y} p(X = x, Y = y) = \sum_{\text{all } y} p(x, y).$$

- The marginal probability distribution of a random variable X is the same mathematical object as what in Lecture 3 we called simply the probability distribution of X .
- Given that we can compute the probability distribution of X and the probability distribution of Y from their joint probability distribution, this joint probability distribution contains at least as much information as these two marginal distributions.

Marginal Probability Distribution

- In fact, the joint probability distribution of the vector (X, Y) contains strictly more information than the sum of the information contained in the marginal distribution of X and the marginal distribution of Y . The reason is that the marginal distributions of X and Y contain no information about the degree of co-movement or association between X and Y .
- In our example:

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$	$p_X(x)$
$X = 1$	0.5	2.7	1.1	1.9	2.1	8.3
$X = 2$	4.6	6.0	7.6	4.2	5.7	28.1
$X = 3$	9.1	4.8	8.1	2.7	3.0	27.7
$X = 4$	25.8	2.2	6.1	0.8	1.0	35.9
$p_Y(y)$	40	15.7	22.9	9.6	11.8	

Marginal Probability Distribution

- What is the probability that a randomly selected worker from the population of civilian non-institutional employed workers 25 years old and over has some college education but not a college degree?
- What is the probability that such a worker is employed in an occupation classified as “natural resources, construction, and maintenance” or classified as “production and transportation”?
- If you randomly select a worker from the 2010 US population of civilian non-institutional employed workers 25 years old and over, is it more likely that it is employed in a professional-managerial occupation or that it has a college degree?

DISCRETE RANDOM VARIABLES: CONDITIONAL PROBABILITY DISTRIBUTION

Conditional Probability Distribution

- The **conditional probability distribution function** of a random variable Y given that a random variable X is equal to some particular value x indicates the probability of each possible value in the support of Y given that the outcome of the random variable X is equal to a particular number x .
- The conditional probability distribution function of Y given $X = x$ is defined as the joint probability distribution function evaluated at $X = x$ divided by the marginal probability of X evaluated at $X = x$.

$$p_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P[\{X = x\} \cap \{Y = y\}]}{P(X = x)} = \frac{p(x, y)}{p_X(x)}$$

- In our example, the probability distribution of X given $Y = 1$ is:

X	1	2	3	4
$P(X Y = 1)$	0.012	0.115	0.228	0.645

Example

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$	$p_X(x)$
$X = 1$	0.5	2.7	1.1	1.9	2.1	8.3
$X = 2$	4.6	6.0	7.6	4.2	5.7	28.1
$X = 3$	9.1	4.8	8.1	2.7	3.0	27.7
$X = 4$	25.8	2.2	6.1	0.8	1.0	35.9
$p_Y(y)$	40	15.7	22.9	9.6	11.8	

- Imagine you randomly select a worker from the population of US workers and this worker has less than a high school diploma, what is the probability that her occupation is classified in the construction category?

$$P(Y = 4|X = 1) = \frac{P(X = 1, Y = 4)}{P(X = 1)} = \frac{1.9}{8.3} = 22.9$$

Example

	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$	$Y = 5$	$p_{\mathbf{x}}(\mathbf{x})$
$X = 1$	0.5	2.7	1.1	1.9	2.1	8.3
$X = 2$	4.6	6.0	7.6	4.2	5.7	28.1
$X = 3$	9.1	4.8	8.1	2.7	3.0	27.7
$X = 4$	25.8	2.2	6.1	0.8	1.0	35.9
$p_{\mathbf{Y}}(\mathbf{y})$	40	15.7	22.9	9.6	11.8	

- Imagine that a randomly selected worker is employed in the construction category, what is the probability that she has less than a high school diploma?

$$P(X = 1 | Y = 4) = \frac{P(X = 1, Y = 4)}{P(Y = 4)} = \frac{1.9}{9.6} = 19.8$$

DISCRETE RANDOM VARIABLES: SIMPSON'S PARADOX

Simpson's Paradox

- Is the application of death penalty racially motivated?
- In 1978, Warren McClesky, a black man, was convicted of killing a white police officer and was sentenced to death in Georgia. In an appeal before the U.S. Supreme Court, McClesky's lawyers argued that imposition of death penalty in Georgia was racially biased against black suspects.
- In 1981, Radelet studied data on the sentence (presence or absence of death penalty) applied to 326 defendants convicted of murder.
- Radelet's data shows:

Race of Defendant	Death Penalty		% Yes
	Yes	No	
White	19	141	$19/160 = 0.118$
Black	17	149	$17/166 = 0.102$
Total	36	290	326

- White defendants are marginally more likely to be sentenced to death penalty.

Simpson's Paradox

- If we compute the joint distribution of “race of defendant” and “death penalty” differentiating by the race of the victim, we obtain:
- For white victims:

Race of Defendant	Death Penalty		% Yes
	Yes	No	
White	19	132	$19/151 = 0.126$
Black	11	52	$11/63 = 0.175$
Total	30	184	214

- For black victims:

Race of Defendant	Death Penalty		% Yes
	Yes	No	
White	0	9	$0/9 = 0$
Black	6	97	$6/103 = 0.058$
Total	6	106	112

Simpson's Paradox

- We now see that blacks are more likely to be sentenced to the death penalty both when the victim is white and when the victim is black.
- Why do the results change when we ignore the race of the victim?
- In order to understand what is going on, it is helpful to look at the joint distribution of the race of the defendant and that of the victim:

Race of Defendant	Race of Victim		% W_v
	White	Black	
White	151	9	$151/160 = 0.944$
Black	63	103	$63/166 = 0.379$

White defendants are more likely to be accused of crimes against white victims, and black defendants against black victims. The defendant being white is associated with the victim being white.

Simpson's Paradox

- It is also helpful to look at the joint distribution of the race of the victim and the probability that the defendant is condemned to the death penalty.

Race of Victim	Death Penalty		% Yes
	Yes	No	
White	30	184	$30/214 = 0.140$
Black	6	106	$6/112 = 0.054$

The probability that the defendant is condemned to the death penalty is much larger if the victim is white than if she is black. The victim being white is associated with the defendant being condemned to the death penalty.

Simpson's Paradox

- Intuition. Even though, both for white and black victims, black defendants are more likely to be condemned, in the aggregate, the fact that
 - white defendants are more likely to be accused of crimes against white victims,
 - any defendant is more likely to be condemned when the victim is white,generates a composition effect such that white defendants are more likely to be condemned overall.
- The positive association between the defendant being white and the defendant being condemned to the death penalty is explained by the presence of an *omitted* variable or *lurking* variable.
- This omitted variable is the race of the victim. The victim being white is a variable that is positively associated with the defendant being white and with the defendant being condemned.

Simpson's Paradox



Figure: Causal Relationship

- The positive association between the defendant being white and the defendant being condemned to the death penalty does not reflect a causal relationship (i.e. discrimination against white defendants).

Simpson's Paradox

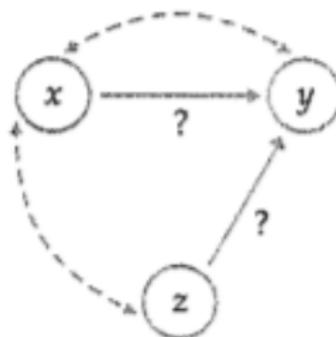


Figure: Confounding Variable

- The positive association between the defendant being white and the defendant being condemned to the death penalty reflects two positive associations:
 - (a) between the defendant being white and the victim being white; and
 - (b) between the victim being white and the defendant being condemned.

Simpson's Paradox

- A more formal explanation using the tools we have learned so far:

$$\begin{aligned} P(\text{Yes}|B_d) &= P(\text{Yes} \cap W_v|B_d) + P(\text{Yes} \cap B_v|B_d) \\ &= P(\text{Yes}|W_v, B_d)P(W_v|B_d) + P(\text{Yes}|B_v, B_d)P(B_v|B_d) \\ &= 0.175 \times 0.379 + 0.058 \times (1 - 0.379) = 0.102 \end{aligned}$$

and

$$\begin{aligned} P(\text{Yes}|W_d) &= P(\text{Yes} \cap W_v|W_d) + P(\text{Yes} \cap B_v|W_d) \\ &= P(\text{Yes}|W_v, W_d)P(W_v|W_d) + P(\text{Yes}|B_v, W_d)P(B_v|W_d) \\ &= 0.126 \times 0.944 + 0 \times (1 - 0.944) = 0.118 \end{aligned}$$

DISCRETE RANDOM VARIABLES: INDEPENDENCE OF RANDOM VARIABLES

[http://www.npr.org/blogs/money/2014/09/19/349650496/
episode-570-the-fine-print](http://www.npr.org/blogs/money/2014/09/19/349650496/episode-570-the-fine-print)

Example: The Fine Print

- Things covered by home insurance
 - *damage caused by riot or commotion*
 - *damage caused by a non-nuclear missile hitting the house (except in the context of war)*
 - *damage caused by lava spilled from a volcano*
 - *fire*
- Things not covered by home insurance
 - *damage caused during a war (including civil war)*
 - *damaged caused by a nuclear missile*
 - *damage caused earthquake generated by a volcano*
 - *damage caused by flooding*
- What is the key difference between the random phenomena that are covered and those that are not covered?

"If you are insuring London in 1939 and war is covered, you are essentially about to have to pay to rebuild all of London."

Example: The Fine Print

- An insurance company is providing insurance to two houses: A and B .
- Let's define the following dummy random variables

$X_A = 1$ if house of client A is hit by a missile not in the context of war,

$Y_A = 1$ if house of client A is hit by a missile in the context of war,

$Z_A = 1$ if house of client A is damaged by a riot.

- Any of these events destroys the house.
- Both houses A and B have equal value: \$100,000.
- Let's assume

$$P(X_A = 1) = P(Y_A = 1) = P(Z_A = 1) = 0.5$$

$$P(X_B = 1) = P(Y_B = 1) = P(Z_B = 1) = 0.5$$

- The insurance company only has \$100,000 to pay in indemnities. If it has to pay more than \$100,000 in indemnities, it will go bankrupt.

Example: The Fine Print

- Let's assume the following joint distributions across the two clients of the insurance company.

		$X_A = 0$	$X_A = 1$
X_B	$X_B = 0$	1/4	1/4
	$X_B = 1$	1/4	1/4

		$Y_A = 0$	$Y_A = 1$
Y_B	$Y_B = 0$	1/2	0
	$Y_B = 1$	0	1/2

		$Z_A = 0$	$Z_A = 1$
Z_B	$Z_B = 0$	0	1/2
	$Z_B = 1$	1/2	0

Example: The Fine Print

- What is the probability that the insurance company goes bankrupt?
- If the insurance company offers insurance against the risk of being hit by a missile not in the context of war:

$$P(\{X_A = 1\} \cap \{X_B = 1\}) = 1/4.$$

- If the insurance company offers insurance against the risk of being hit by a missile in the context of war:

$$P(\{Y_A = 1\} \cap \{Y_B = 1\}) = 1/2.$$

- If the insurance company offers insurance against the risk of being damaged by a riot:

$$P(\{Z_A = 1\} \cap \{Z_B = 1\}) = 0.$$

Example: The Fine Print

- Therefore, an insurance company which is trying to avoid going bankrupt is more likely to offer coverage, in this order, for
 - ① damage due to riot (Z),
 - ② damage due to a missile not in the context of war (X),
 - ③ damage due to a missile in the context of war (Y).
- Notice that the marginal probability distribution of the random variables X_A , X_B , Y_A , Y_B , Z_A and Z_B is identical.
- The only thing that differentiates X , Y , and Z is the degree of dependence between X_A and X_B , Y_A and Y_B , and Z_A and Z_B .
 - X_A and X_B are **independent**
 - Y_A and Y_B are positively associated.
 - Z_A and Z_B are negatively associated.

Independence

- Two discrete random variables X and Y are **independent** if the distribution of X conditional on Y is the same as the marginal distribution of X .
- More formally, two discrete random variables X and Y are said to be **independent** if, and only if, for **all** possible values of y and x , it holds:

$$P(Y = y|X = x) = P(Y = y).$$

- There are several equivalent ways of stating the definition of independence

$$P(Y = y|X = x) = P(Y = y)$$

$$P(X = x|Y = y) = P(X = x)$$

$$P(\{X = x\} \cap \{Y = y\}) = P(Y = y)P(X = x)$$

$$F(Y = y|X = x) = F(Y = y)$$

$$F(X = x|Y = y) = F(X = x)$$

$$F(\{X = x\} \cap \{Y = y\}) = F(Y = y)F(X = x)$$

Independence

- Are the following two variables independent?

		$Y = 0$	$Y = 1$
		1/6	1/6
$X = 0$	1/6	1/6	
	2/6	2/6	

Assuming Events are Independent When They are Not

- Assuming events are independent when they are not is a mistake that can have very important consequences.
- **Example 1** In the 1990s, British prosecutors committed a grave miscarriage of justice because of an improper use of probability.
- The mistake arose in the context of sudden infant death syndrome (SIDS), a phenomenon in which a perfectly healthy infant dies in his or her crib.
- SIDS was (and is still) a medical mystery.
- Because these infant deaths were so poorly understood, they bred suspicion about possible parental negligence.
- British prosecutors and courts became convinced that one way to separate parental negligence from natural deaths would be to focus on families in which there were multiple sudden deaths.

Assuming Events are Independent When They are Not

- Sir Roy Meadow, a prominent British pediatrician, was a frequent expert witness on this point. He claimed that the incidence of SIDS is 1 in 8,500 and that the chance of having two events of SIDS in the same family would be

$$\frac{1}{8,500} \times \frac{1}{8,500} = \frac{1}{73,000,000}$$

- Based on this calculation, juries decided that two events of SIDS in one couple was evidence of parental negligence, and sent many parents to prison on the basis of this testimony on the statistics of SIDS.
- The Royal Statistical Society pointed out the flaw in the reasoning. It is quite possible that there is a link –something genetic, for instance– which would make a family that had suffered one event of SIDS more likely to suffer another.
- In 2004, the British government announced that it would review 258 trials in which parents had been convinced of murdering their infant children.

Assuming Events are Independent When They are Not

- **Example 2.** The collapse of AIG was due to an once-obscure instrument known as credit default swap (CDS). The CDS is an insurance on bonds.
- Imagine a large bank buys some bonds issued by General Electric. The bank expects to receive a steady stream of payments from GE over the years.
- But the bank might think there is a chance that GE might go bankrupt. If this happens, the bank does not get any more of these payments.
- To insure itself, the bank might decide to buy a CDS. If GE never goes bankrupt, the bank has lost in net the price of the CDS. If GE goes bankrupt, the bank gets the remaining bond payments from whoever sold the CDS.
- AIG sold CDSs that covered more than \$440 billion in bonds. AIG estimated there was a very low probability that companies like GE would default on many of these bonds **at the same time**. However, once some bonds start defaulting, other bonds are more likely to default.
- As bonds started to default, AIG did not have enough money to cover all the bonds that were defaulting.

DISCRETE RANDOM VARIABLES: REVIEW OF LECTURE 5

Homework

- Consider a pair of random variables
 - X : years of education.
 - Y : annual earnings in thousands of dollars.

with the following joint probability distribution

		$X = 8$	$X = 12$	$X = 16$
Y	25	0.15	0.15	0
	45	0.05	0.15	0.1
	70	0	0.15	0.25

- Compute
 - $F(12,45)$
 - The marginal probability distribution of X and Y .
 - The conditional probability distribution of X given $Y = 45$.
 - The conditional probability distribution of Y given $X = 12$
 - Are these two variables independent?

WWS 507c: Quantitative Analysis

Lecture 6: Random Variables and Probability Distributions (cont.)

Princeton University

September 30, 2014

INTRODUCTION

Introduction

- In lecture 5, we introduced the concepts of
 - joint probability distribution,
 - joint cumulative distribution,
 - conditional probability distribution,
 - independence,

for the case of discrete random variables.

- In this lecture, we will
 - First: review what we learned about joint distributions of discrete variables.
 - Second: learn that, in the case of bivariate distributions (i.e. distributions of **two** variables) we can represent their joint probability distribution and joint cumulative distribution graphically.
 - Third: learn how to apply the concepts of joint probability distribution, joint cumulative distribution, conditional probability distribution and independence to the case of continuous random variables.

DISCRETE RANDOM VARIABLES: REVIEW OF LECTURE 5 AND GRAPHICAL REPRESENTATION

Example

- Consider a pair of random variables
 - X : years of education.
 - Y : annual earnings in thousands of dollars.

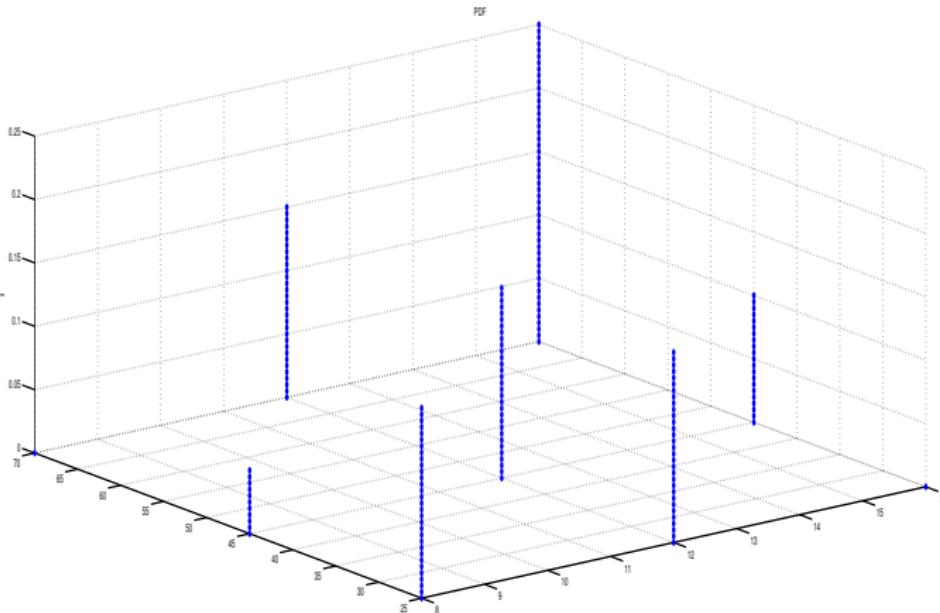
with the following joint probability distribution

		$X = 8$	$X = 12$	$X = 16$
Y	25	0.15	0.15	0
	45	0.05	0.15	0.1
	70	0	0.15	0.25

- Review questions. Compute:
 - $F(12,45)$
 - The marginal probability distribution of X and Y .
 - The conditional probability distribution of X given $Y = 45$.
 - $P(\{Y \leq 45\} \cap \{X \geq 12\})$ and $P(\{Y \leq 45\} \cap \{X \geq 12\} | X = 16)$.
 - Are these two random variables independent?

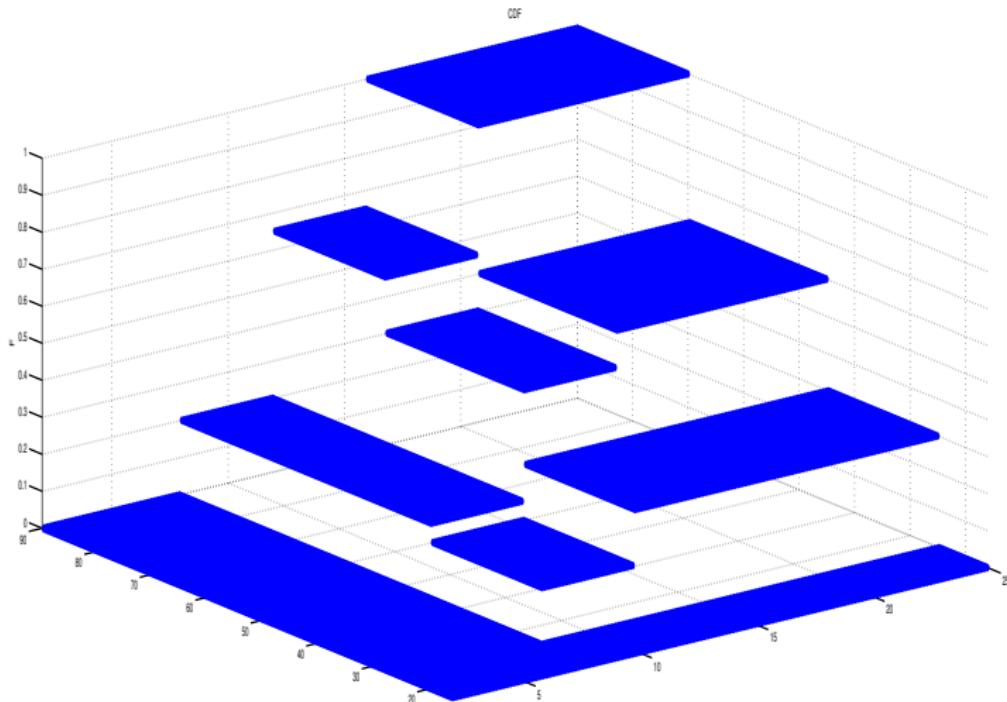
Example: Graphical Representation of PDF

- The vertical axis captures the value of the PDF.
- The points of the (X, Y) space with a blue spike represent the support of the random vector (X, Y) .



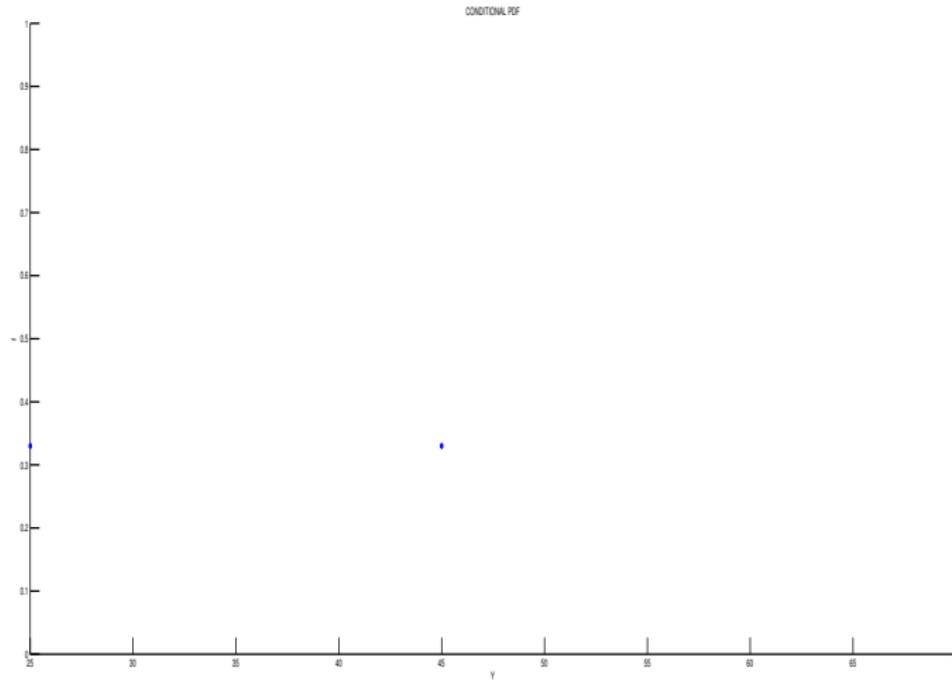
Example: Graphical Representation of CDF

- The vertical axis captures the value of the CDF.



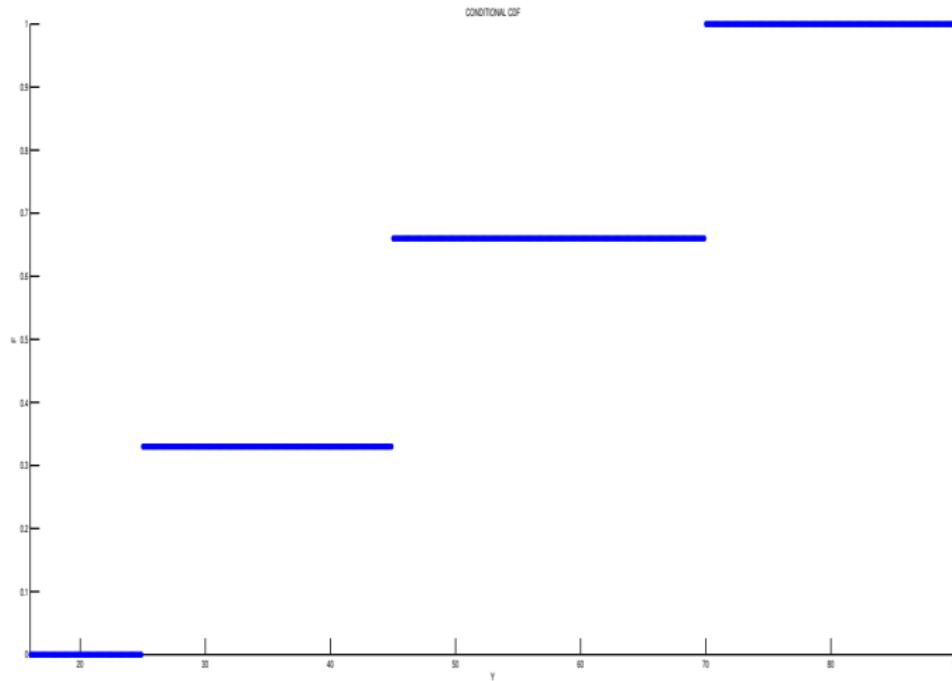
Example: Graphical Representation of Conditional PDF

- We can also represent the conditional probability $P(Y|X = 12)$



Example: Graphical Representation of Conditional CDF

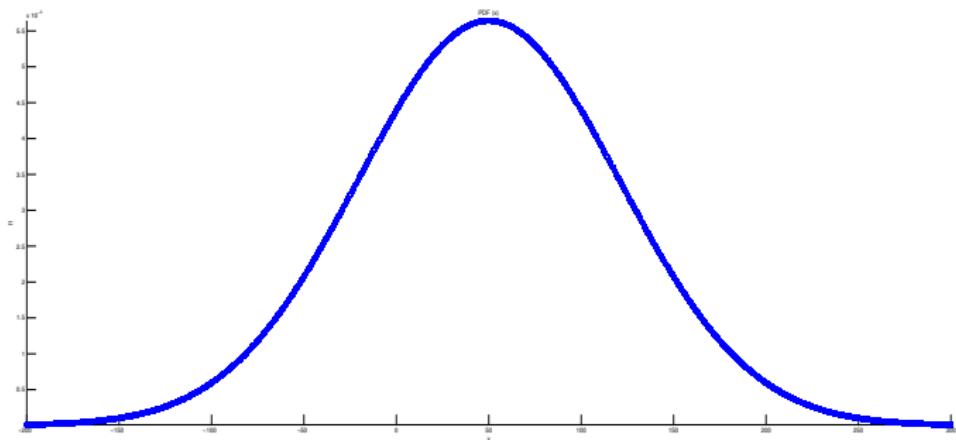
- We can also represent the conditional distribution $F(Y|X = 12)$



JOINT DISTRIBUTION OF CONTINUOUS RANDOM VARIABLES

Joint Distribution of Continuous Random Variables

- Planning for college expenses is one of the biggest financial projects that a family can undertake. The Smiths have to pay \$100,000 in tuition for their child. In order to obtain this money, the Smiths has decided to go to the casino. Imagine that both Mr. and Mrs. Smith decide to play roulette together. Let's denote by X the random variable that captures the money they have as they get out of the casino. Its probability density function is:

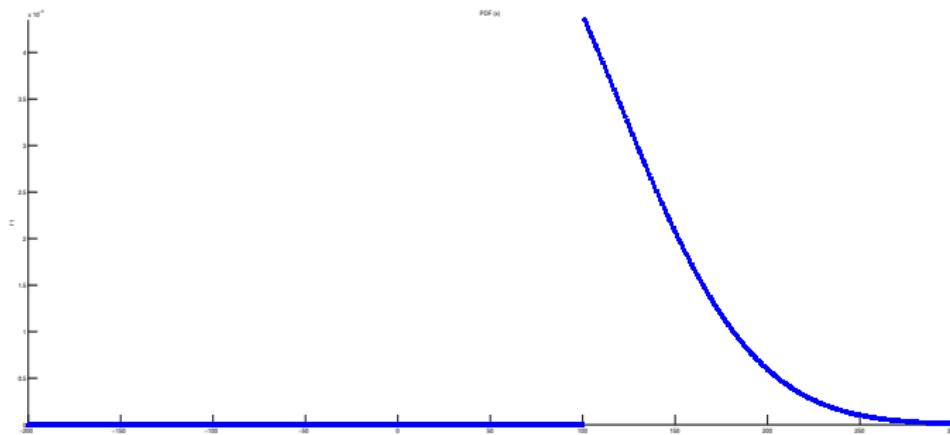


Joint Distribution of Continuous Random Variables

- Note that the support that the random variable X is $(-\infty, \infty)$.
- The probability that the returns to the trip to the casino will be enough to pay for college is:

$$P(X \geq 100) = \int_{100}^{\infty} f(x)dx = 1 - F(100)$$

or, equivalently, the area under the following function



Joint Distribution of Continuous Random Variables

- Imagine now that Mr. and Mrs. Smith decide to play separately. Let's use X to denote the money that Mr. Smith will earn and Y to denote the money that Mrs. Smith will earn.
- How would you compute the probability that the Smiths gets out of the casino with enough money to pay their child's college tuition?
- Mathematically, this question asks for $P(X + Y \geq 100)$.
- There are two ways to compute this probability
 - either we send the Smiths to the casino for many many times and we compute the fraction of times that their joint gains are above \$100,000...
 - or we know the joint probability density function of X and Y , $f(X, Y)$.

Joint Distribution of Continuous Random Variables

- The **joint probability density function** of a random vector (X, Y) is defined as a function such that, for any subset A of the support of (X, Y) (i.e. for any set A of values that (X, Y) might take), the probability

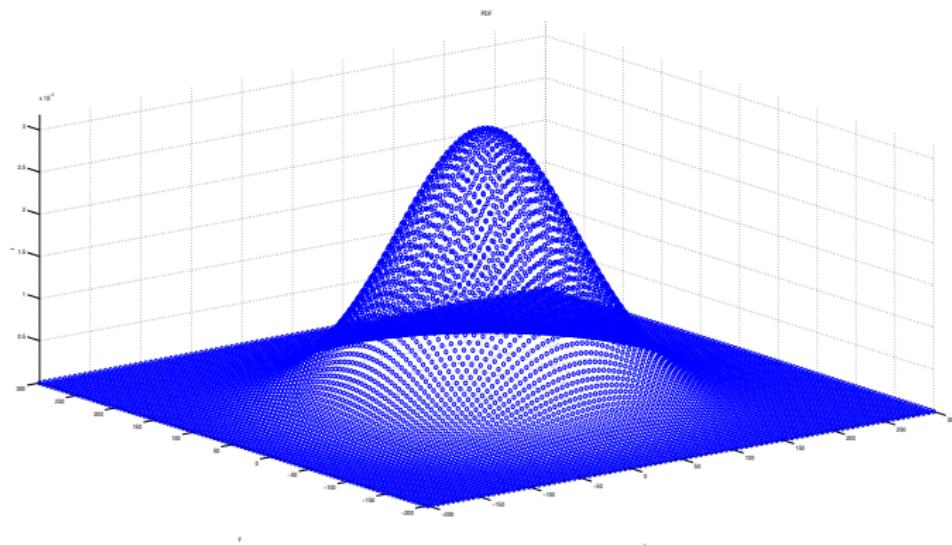
$$P((X, Y) \in A)$$

is equal to the **volume** under the joint probability density function of (X, Y) and above the area A .

- Let's denote the joint probability density function of the random vector (X, Y) as $f(X, Y)$.
- Therefore, $P(X + Y \geq 100)$ is equal to the **volume** under $f(X, Y)$ and above those points (x, y) in the support of (X, Y) such that $x + y > 100$.

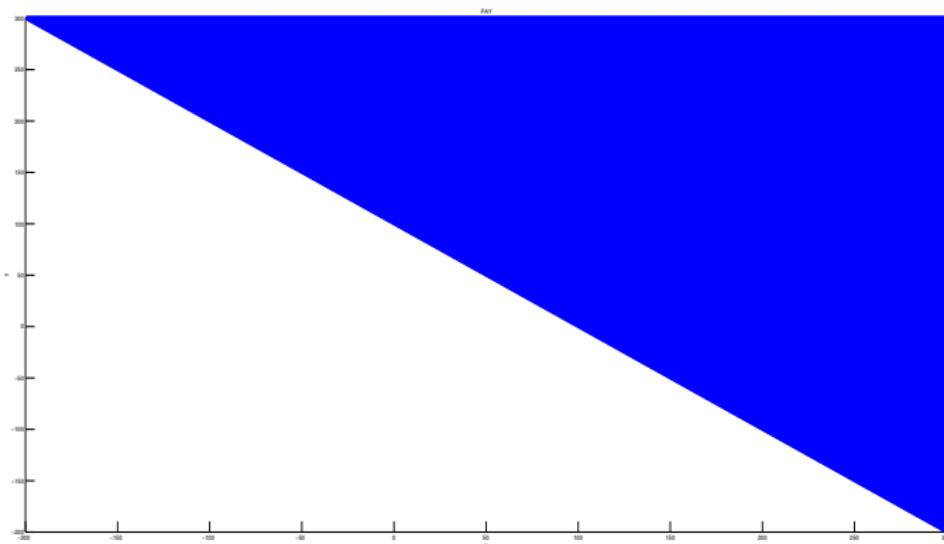
Joint Distribution of Continuous Random Variables

- Assume that the joint probability density function of the random vector (X, Y) is



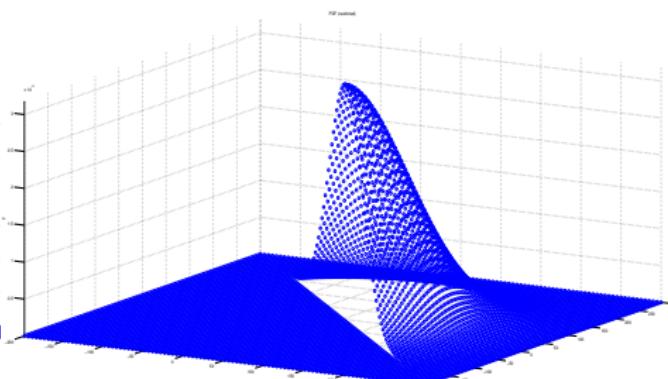
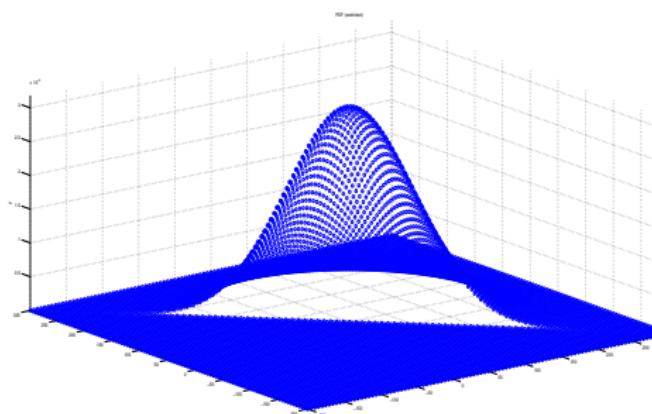
Joint Distribution of Continuous Random Variables

- The support of (X, Y) is: $(-\infty, \infty) \times (-\infty, \infty)$
- However, the only outcomes that would allow the Smiths to pay for their child's college are...



Joint Distribution of Continuous Random Variables

- Therefore, $P(X + Y \geq 100)$ is only the area under the following function...



Joint Distribution of Continuous Random Variables

- How do we compute the volume under a joint density function?
- We use integrals.

$$P((X, Y) \in A) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}\{(x, y) \in A\} f(x, y) dx dy$$

where $\mathbb{1}\{(X, Y) \in A\}$ is a function that takes value 1 if (x, y) is in the set A , and 0 if it is not.

- Therefore,

$$P(X + Y \geq 100) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}\{x + y \geq 100\} f(x, y) dx dy$$

- How do we compute this integral?

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}\{x + y \geq 100\} f(x, y) dx dy &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}\{x \geq 100 - y\} f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} \int_{100-y}^{\infty} f(x, y) dx dy \end{aligned}$$

Joint Distribution of Continuous Random Variables

- Example. Let (X, Y) be uniform on the unit square. Then,

$$f(x, y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and

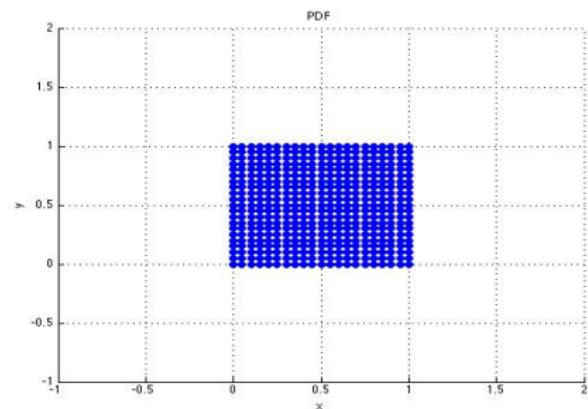
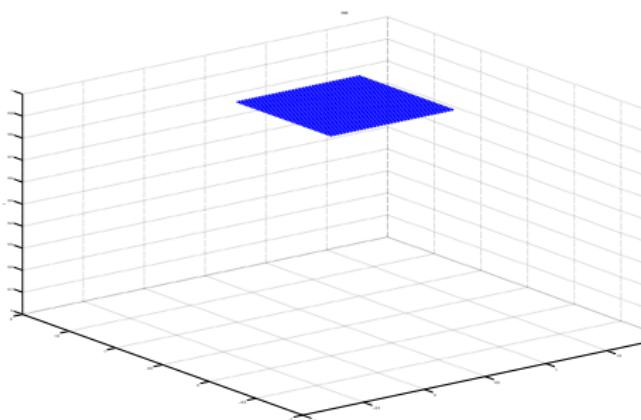
$$\begin{aligned} P(X < 1/2, Y < 1/2) &= \int_0^{1/2} \int_0^{1/2} f(x, y) dx dy = \int_0^{1/2} \int_0^{1/2} 1 dx dy = \\ &= \int_0^{1/2} \left[x \right]_0^{1/2} dy = \int_0^{1/2} \frac{1}{2} dy = \frac{1}{2} \int_0^{1/2} 1 dy = \frac{1}{4}. \end{aligned}$$

Joint Distribution of Continuous Random Variables

- Visually, the joint probability distribution

$$f(x, y) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

is



Continuous bivariate distributions: PDF

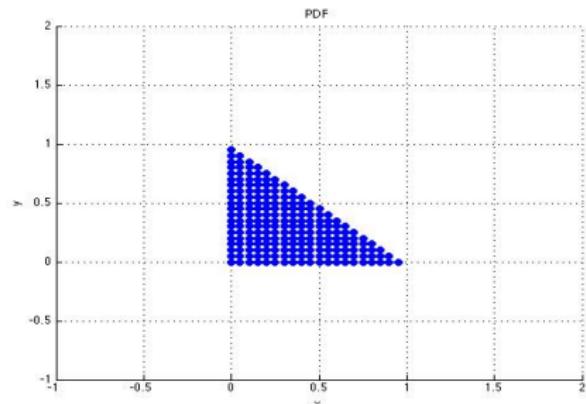
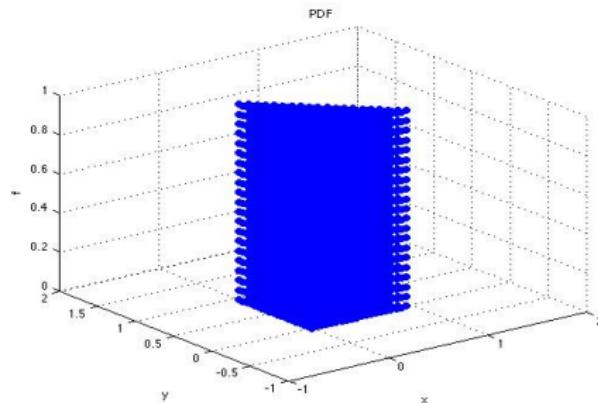
- More examples (this is quite tricky!)

$$\begin{aligned} P(X + Y \leq 1) &= \int_0^1 \int_0^{\min\{1-y, 1\}} 1 dx dy = \int_0^1 \int_0^{1-y} 1 dx dy = \int_0^1 \left[x \right]_0^{1-y} dy = \\ &= \int_0^1 (1-y) dy = \left[y - \frac{y^2}{2} \right]_0^1 = 1 - \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

$$\begin{aligned} P(X + Y \leq 1.5) &= \int_0^1 \int_0^{\min\{1.5-y, 1\}} 1 dx dy = \\ &= \int_0^{0.5} \int_0^{\min\{1.5-y, 1\}} 1 dx dy + \int_{0.5}^1 \int_0^{\min\{1.5-y, 1\}} 1 dx dy = \\ &= \int_0^{0.5} \int_0^1 1 dx dy + \int_{0.5}^1 \int_0^{1.5-y} 1 dx dy = \\ &= \frac{1}{2} + \int_{0.5}^1 (1.5-y) dy = \frac{1}{2} + \left[1.5y - \frac{y^2}{2} \right]_{0.5}^1 = 0.875. \end{aligned}$$

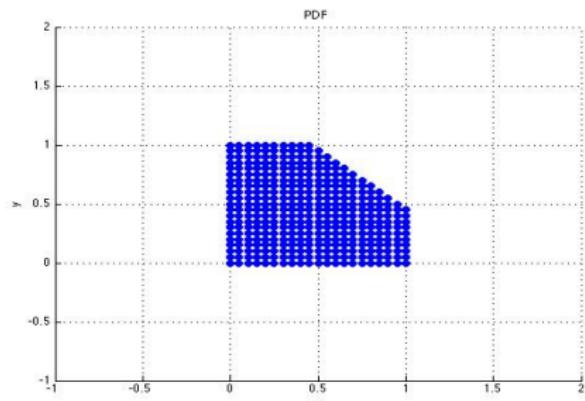
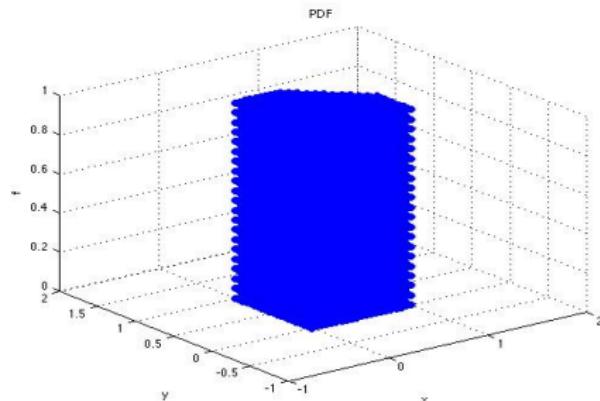
Continuous bivariate distributions: PDF

- When computing $P(X + Y \leq 1)$, we are actually computing the volume of the left figure...



Continuous bivariate distributions: PDF

- and computing $P(X + Y \leq 1.5)$ is actually equivalent to computing the volume of this left figure



CONTINUOUS RANDOM VARIABLES: JOINT CUMULATIVE DISTRIBUTION FUNCTION

Joint Cumulative Distribution Function

- We can compute the joint cumulative distribution function (CDF) using the joint probability density function:

$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(X, Y) dXdY,$$

where the joint CDF evaluated at a particular point (x, y) indicates the probability that X takes a value smaller or equal to x **and** Y takes a value smaller or equal to y

$$F(x, y) = P(X \leq x, Y \leq y)$$

- We can go from the joint PDF to the joint CDF by computing integrals. Conversely, we can go from the joint CDF to the joint PDF by computing derivatives

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

Joint Cumulative Distribution Function

- Example. Consider a pair of random variables (X, Y) with joint PDF

$$f(x, y) = \begin{cases} 4xy & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- The joint CDF for (x, y) such that $0 \leq x \leq 1, 0 \leq y \leq 1$ is

$$\begin{aligned} F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(X, Y) dY dX = \int_0^x \int_0^y 4XY dY dX = 4 \int_0^x X \left[\int_0^y Y dY \right] dX = \\ &= 4 \int_0^x X \frac{y^2}{2} dY = 2y^2 \int_0^x X dX = y^2 x^2. \end{aligned}$$

- The joint CDF for (x, y) such that $0 \leq x \leq 1, 1 \leq y$ is

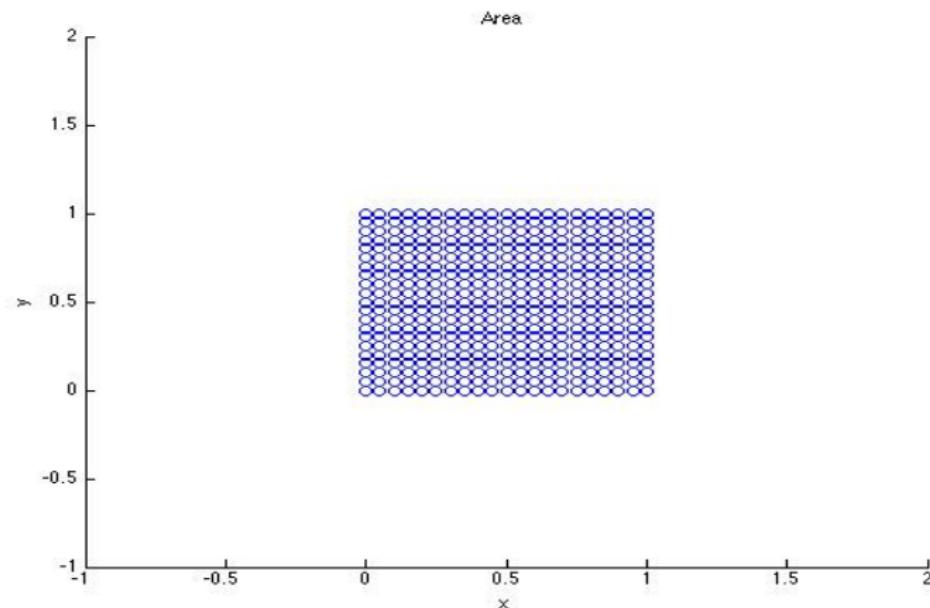
$$\begin{aligned} F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(X, Y) dY dX = \int_0^x \int_0^1 4XY dY dX = 4 \int_0^x X \left[\int_0^1 Y dY \right] dX = \\ &= 4 \int_0^x X \frac{1^2}{2} dX = 2 \int_0^x X dX = x^2. \end{aligned}$$

Joint Cumulative Distribution Function

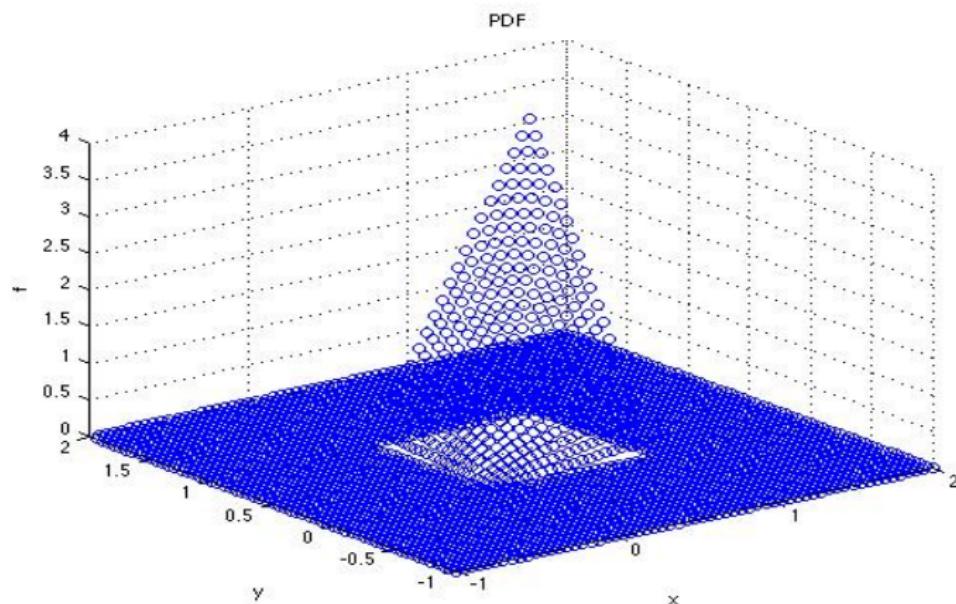
- Therefore, the joint CDF is

$$F(x, y) = \begin{cases} 0 & \text{if } x < 0 \text{ or } y < 0 \\ y^2x^2 & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ x^2 & \text{if } 0 \leq x \leq 1, 1 < y, \\ y^2 & \text{if } 1 < x, 0 \leq y \leq 1, \\ 1 & \text{if } 1 < x, 1 < y, \end{cases}$$

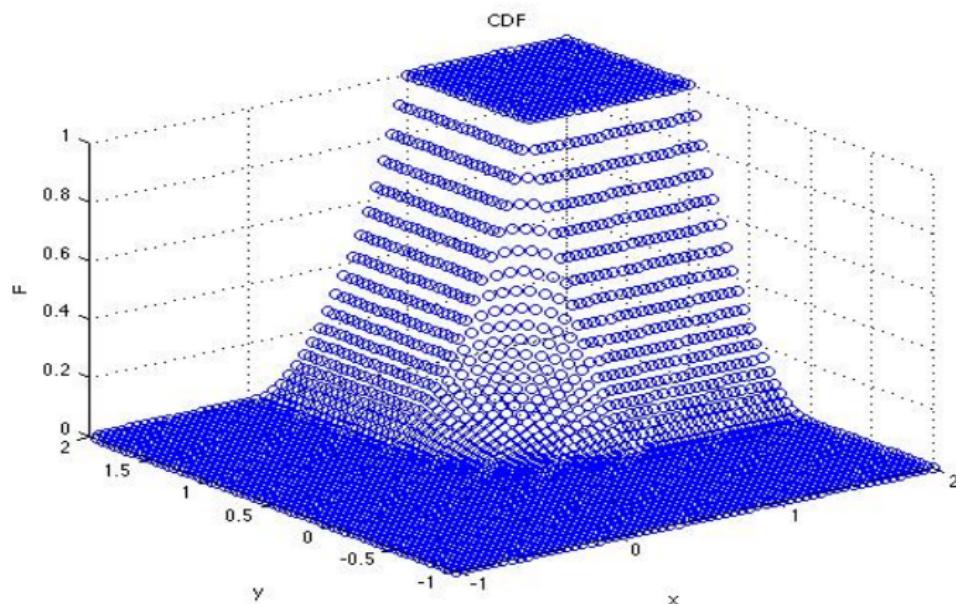
Joint Cumulative Distribution Function



Joint Cumulative Distribution Function



Joint Cumulative Distribution Function



CONTINUOUS RANDOM VARIABLES: MARGINAL PROBABILITY DENSITY FUNCTION

Marginal Density Function

- We can compute the marginal density function of a random variable X from the joint probability density function of X with any other random variable Y as

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$

- If

$$f(x, y) = \begin{cases} 4xy & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

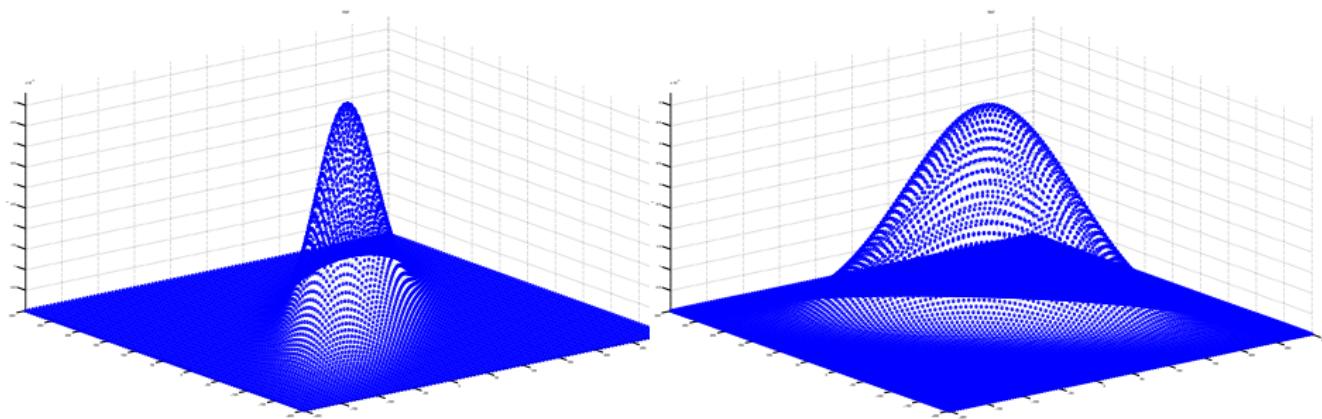
then

$$f_X(x) = \int_0^1 4xy \mathbb{1}\{0 \leq x \leq 1\} dy = 4x \mathbb{1}\{0 \leq x \leq 1\} \int_0^1 y dy = 2x \mathbb{1}\{0 \leq x \leq 1\}$$

$$f_Y(y) = \int_0^1 4xy \mathbb{1}\{0 \leq y \leq 1\} dx = 4y \mathbb{1}\{0 \leq y \leq 1\} \int_0^1 x dx = 2y \mathbb{1}\{0 \leq y \leq 1\}$$

Marginal Density Function

- Knowledge of the marginal probability density function of two random variables X and Y is not enough to recover their joint probability density function



In the case of these two joint probability distributions, the marginal distributions of X and Y are identical! (and also identical to the marginal distributions of X and Y behind the joint distribution in slide 15!)

CONTINUOUS RANDOM VARIABLES: CONDITIONAL PROBABILITY DENSITY FUNCTION

Conditional Density Function

- The conditional probability density function of some variable Y given some other variable X is defined as the joint probability density function of the vector (X, Y) divided by the marginal distribution of X :

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}.$$

- If

$$f(x, y) = \begin{cases} 4xy & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

then

$$f_{Y|X}(y|x) = \frac{4xy \mathbb{1}\{0 \leq x \leq 1\} \mathbb{1}\{0 \leq y \leq 1\}}{2x \mathbb{1}\{0 \leq x \leq 1\}} = 2y \mathbb{1}\{0 \leq y \leq 1\},$$

$$f_{X|Y}(x|y) = \frac{4xy \mathbb{1}\{0 \leq x \leq 1\} \mathbb{1}\{0 \leq y \leq 1\}}{2y \mathbb{1}\{0 \leq y \leq 1\}} = 2x \mathbb{1}\{0 \leq x \leq 1\}.$$

Independence

- Two random variables X and Y are said to be independent if

$$P(Y \leq y | X = x) = P(Y \leq y),$$

for all possible values of y and x .

- There are several equivalent ways of stating the definition of independence

$$f_{Y|X}(y|x) = f_Y(y)$$

$$F_{Y|X}(y|x) = F_Y(y)$$

$$f_{X|Y}(x|y) = f_X(x)$$

$$F_{X|Y}(x|y) = F_X(x)$$

$$f(x, y) = f(x)f(y)$$

$$F(x, y) = F(x)F(y)$$

and these equalities should hold for every value of x and y .

- Check these definitions with the example in the previous slide. Are X and Y independent?

More than two variables

- The results that have been presented above for two random variables, X and Y , can be extended to more than two variables.
- Consider a set of n random variables Y_1, Y_2, \dots, Y_n .
- The joint CDF is

$$F(y_1, y_2, \dots, y_n) = P[(Y_1 \leq y_1) \cap (Y_2 \leq y_2) \cap \dots \cap (Y_n \leq y_n)].$$

and the joint PDF is

$$f(y_1, y_2, \dots, y_n) = \frac{\partial^n F(y_1, y_2, \dots, y_n)}{\partial y_1 \partial y_2 \dots \partial y_n}.$$

- Analogously, we can define a marginal distribution

$$f(y_1, y_2, \dots, y_s) = \int_{\text{all } y_{s+1}} \dots \int_{\text{all } y_n} f(y_1, \dots, y_s, y_{s+1}, \dots, y_n) dy_{s+1} dy_{s+2} \dots dy_n$$

and a conditional distribution

$$f(y_{s+1}, \dots, y_n | y_1, \dots, y_s) = \frac{f(y_1, \dots, y_s, y_{s+1}, \dots, y_n)}{f(y_1, y_2, \dots, y_s)}$$

WWS 507c: Quantitative Analysis

Lecture 7: Moments of a Distribution

Princeton University

October 2, 2014 + October 7, 2014

INTRODUCTION

Introduction

- Imagine I was interested in obtaining information about $X = \text{SAT verbal scores in 2013 of students born in NJ.}$
- Imagine also that you were to have access to data on the SAT verbal scores for every student taking the exam in 2013.
- How would you transmit to me information about X ?
- Imagine that I was also interested in comparing the random variable X to the random variable

$Y = \text{SAT verbal scores in 2013 of students born in Michigan}$

- Specifically, imagine that I was looking for an answer to the question:
 - Did students from NJ obtain better grades than students from Michigan?
- How would you use the information you have to answer my question?

Introduction

- One option is that you use what we have learned in class so far and that you give me the PDF of X and the PDF of Y .
- The PDF of X would look something like...

X	$P(X)$
600	0.0003
601	0.0001
602	0.0002
603	0.0004
\vdots	

and you could construct a similar PDF for the random variable Y .

Introduction

- However, knowing the PDF of X and Y would not be very useful in this case.
- The contain too many numbers...too much information.
- The PDFs are useful to compute probabilities; e.g.

$$P(1000 < X < 2000)$$

$$P(Y > 1500)$$

but, if you want to get a quick idea of the distribution of a random variable, knowing its probability distribution is likely to contain so much information that, in the end, you do not learn much.

- Imagine that you were interested in learning something about the income distribution of households in the U.S. and I were to tell you the fraction of U.S. households that have each possible income: fraction of households with \$1 USD, fraction of households with \$2 USD...

Introduction

- We need ways to summarize the information that is contained in these probability distribution functions into a few numbers.
- The numbers we will use to summarize a probability distribution function are the so-called **moments** of a distribution.
- Specifically, in order to summarize univariate probability distributions, we will learn four moments:
 - mean (also known as expectation)
 - variance
 - skewness
 - kurtosis
- In order to summarize the dependency or association between two random variables X and Y , we will learn two measures of dependency:
 - covariance
 - correlation coefficient

MOMENTS OF DISTRIBUTION OF SCALAR RANDOM VARIABLE: MEAN OR EXPECTATION

Mean

- Imagine that you open the newspaper and you read that the life expectancy at birth in the US in 2014 is 79.56 years.
- What does this mean?
- Let me give you a clue. The definition of life expectancy at birth in the US in 2014 is the *mean number of years to be lived by children born in the US in 2014, if mortality at each age remains constant in the future.*
- What does this mean?
- It means that, if nothing changes in the world (no health discoveries, no increase pollution, etc), and we were to compute the average age at death of all babies born in the US in 2014, that average would be 79.56.
- **Reminder.** The **average** of a sequence of N numbers x_1, x_2, \dots, x_N is

$$\frac{1}{N} \sum_{i=1}^N x_i.$$

Mean

- The **mean** of a random variable X is the average outcome that we would obtain if we were to repeat the random phenomenon on which X is defined many many many times.
- Imagine the random variable

X = number obtained if we were to roll a 6-sided die

- Imagine a different random variable

$Y = 1$ if we obtain heads when tossing a fair coin

$Y = 0$ if we obtain tails when tossing a fair coin

- Imagine a third variable

Z = yearly wage of a randomly selected worker in the U.S.

- How would you use the definition of **mean** from the first bullet point to compute the mean of X , Y , and Z ?

Mean

- Luckily, in order to compute the mean of a random variable, we actually do not need to observe many many outcomes from that random variable.
- The probability distribution function (for discrete random variable) or the probability density function (for continuous random variables) of a random variable X is everything we need to know in order to compute the mean of X .
- How do we use the PDF to compute the mean?
- In the case of discrete random variables:

$$\mu_X = \sum_{\text{all } x} xP(X = x) = \sum_{\text{all } x} xp(x)$$

- In the case of continuous random variables:

$$\mu_X = \int_{-\infty}^{\infty} xf(x)dx$$

- It is common to use the greek letter μ to denote the mean of a distribution.

Mean

- Remember that, just as probabilities are an idealized description of long-run proportions, the mean of a probability distribution describes the long-run average outcome.
- The mean of a random variable X is also usually called the **expectation** of X .
- The expectation of X is usually denoted $\mathbb{E}[X]$.
- Therefore, it is common to see, for discrete random variables,

$$\mathbb{E}[X] = \sum_{\text{all } x} xP(X = x) = \sum_{\text{all } x} xp(x)$$

and, for continuous random variables,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Mean: Examples

- Let $X \sim \text{Bernouilli}(p)$,

$$\mu_X = \sum_{x=0}^1 xp(x) = 0 \times (1-p) + 1 \times p = p.$$

- Flip a fair coin two times and denote X the number of heads.

$$\mu_X = \sum_{x=0}^2 xp(x) = (0 \times \frac{1}{4}) + (1 \times \frac{1}{2}) + (2 \times \frac{1}{4}) = 1.$$

- Let $X \sim \text{Uniform}(-1, 3)$,

$$\mu_X = \int_{-1}^3 x \frac{1}{4} dx = \frac{1}{4} \left[\frac{x^2}{2} \right]_{-1}^3 = \frac{1}{8} (9 - 1) = 1.$$

- Roll a die and denote by X the number you obtain. Compute μ_X .

Mean: Examples

- Choose a number between 1 and 6. Once you have chosen the number, I will roll a die. If you chose the number I obtain, I will pay you \$10. If I obtain a different number, I will pay you nothing. Denote by X the random variable capturing the amount of money I will pay you. What is the mean of X ?
- Consider a pair of random variables (X, Y) with joint PDF

$$f(x, y) = \begin{cases} 4xy & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What is the mean of X ?

Mean

- Now we know how to compute the mean or expectation of a random variable X but, why is this useful?
- Imagine that you are planning to drive to White Mountains this weekend and that Google maps offers you two possible routes. The expected length of the trip if you take route 1 is 6 hours and 16 minutes. If you take route 2, the expected length of the trip is 6 hours and 31 minutes. Imagine that you would like to arrive as soon as possible.

Which route shall you take? Why?

- Bear in mind that you are only going to drive once and there is no guarantee that the realization of a random event is identical to its mean!

Mean and Median

- The usefulness of the mean is that it is a **measure of the location of a distribution**.
- If two distributions are identical but one is shifted to the right, then the distribution shifted to the right will have a higher mean.
- But there are other measures of locations of distributions. The most famous alternative to the mean is the **median**.
- The **median** of a random variable X is a number such that, if we were to repeat the random phenomenon on which X is defined many many many times, 50% of the times we would observe an outcome smaller than the median and 50% we would observe an outcome larger than the median.
- If two distributions are identical but one is shifted to the right, then the distribution shifted to the right will also have a higher median.
- In fact, if a distribution is symmetric, then the median and mean are identical.

Mean and Median

- If mean and median are two different measures of location of a distribution, why are they different?
- Mean and median of a random variable X are not only a measures of location of a distribution but they are also, in some way, the **best/closest** possible guess that you could make about the outcome of the random phenomenon generating the random variable X .
- The difference between mean and median is that they use different definitions of distance. They use different definitions of being close.
- The expectation is the number that minimizes the expected (i.e. mean) squared deviation. In other terms, if you had to pick one number, take many many draws from a distribution and compute the square of the difference between each draw and the chosen number, then the number that would minimize the average of these squared differences is the expectation of the distribution.

Mean and Median

- The median is the number that minimizes the expected (i.e. mean) absolute deviation. In other terms, if you had to pick one number, take many many many draws from a distribution and compute the absolute value of the difference between each draw and the chosen number, then the number that would minimize the average of these absolute differences is the expectation of the distribution.
- While the expectation of a random variable solves the problem

$$\min_a \mathbb{E}[X - a]^2,$$

the median of a random variable solves the problem

$$\min_a \mathbb{E}[|X - a|].$$

Mean and Median

- Imagine the following situation:

- I give you the distribution of temperature in Princeton on February 1st;
- I ask you to predict which temperature we will have on February 1st;
- We agree that you will pay to me

$$\$ (\text{actual temperature} - \text{your guess})^2$$

- Example: if you predict $30^{\circ}F$ and the true temperature is $15^{\circ}F$, you will pay to me $\$15^2 = \225
- The guess that will minimize the expected amount of money you will pay is the expectation of the distribution of temperature in Princeton on Feb. 1st.
- If we had agreed that you would pay to me

$$\$ |\text{actual temperature} - \text{your guess}|,$$

then the guess that will minimize the expected amount of money you will pay me is the median of the distribution of temperature.

Expectation of Functions of Random Variables

- If $g(\cdot)$ is a function of X , then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad \mathbb{E}[g(X)] = \sum_{\text{all } x} g(x)p(x)$$

- In the particular case in which we choose $g(X) = \mathbb{1}\{X \in A\}$,

$$\mathbb{E}[g(X)] = \mathbb{E}[\mathbb{1}\{X \in A\}] = \int_{-\infty}^{\infty} \mathbb{1}\{x \in A\}f(x)dx = P(X \in A).$$

| The probability of an event is just the expectation of a dummy variable taking value 1 whenever that event is true.

- If X is such that $P(X = a) = 1$, then $\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = a$.
- If a and b are constants and $g(\cdot)$ is a function of X , then

$$\begin{aligned}\mathbb{E}[a + bg(X)] &= a + \int_{-\infty}^{\infty} bg(x)f(x)dx = a + b \int_{-\infty}^{\infty} g(x)f(x)dx \\ &= a + b\mathbb{E}[g(X)].\end{aligned}$$

Expectation of Functions of Random Variables

- Examples:

- Let $X \sim \text{Uniform}(0, 1)$ and $Y = g(X) = \exp(X)$, then

$$\mathbb{E}[g(X)] = \int_0^1 g(x)f(x)dx = \int_0^1 \exp(x)dx = e - 1.$$

- Take a stick of unit length and break it at random. Let Y be the length of the longer piece. What is the mean of Y ?

If X is the break point, then $X \sim \text{Uniform}(0, 1)$, and $Y = \max\{X, 1 - X\}$.

Then, $Y = 1 - X$ if $0 < X < 1/2$, and $Y = X$ if $1/2 < X < 1$. Hence,

$$\begin{aligned}\mathbb{E}[Y] &= \int_0^1 \max\{x, 1 - x\}f(x)dx = \int_0^1 \max\{x, 1 - x\}dx \\ &= \int_0^{1/2} (1 - x)dx + \int_{1/2}^1 xdx = \left[x - \frac{x^2}{2}\right]_0^{1/2} + \left[\frac{x^2}{2}\right]_{1/2}^1 \\ &= \frac{1}{2} - \frac{1}{8} + \frac{1}{2} - \frac{1}{8} = \frac{3}{4}.\end{aligned}$$

MOMENTS OF DISTRIBUTION OF SCALAR RANDOM VARIABLE: VARIANCE AND STANDARD DEVIATION

Variance/Standard Deviation

- The mean is a measure of the center of a distribution.
- Once we know where the center of a distribution is located, it is immediate to wonder how concentrated the distribution is around that center.
- The variance and the standard deviation give us the answer to this question.
- They are measures of the spread of a distribution **around its mean**.
- Define X as the outcome of rolling a die.
- Can you show that the expectation of X is 3.5?
- Imagine that you were to roll the die many many many times. Denote the outcome of roll i as x_i . After each roll i , you compute

$$y_i = (x_i - 3.5)^2.$$

The variance of the random variable X would tell you what the average of the numbers y_i you are computing would be after having rolled the die many many times.

Variance/Standard Deviation

- Therefore, the variance is $\mathbb{V}(X) = \sigma_X^2 = \mathbb{E}[(X - \mu_X)^2]$, where $\mu_X = \mathbb{E}(X)$
- For continuous random variables, we compute the variance as

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx.$$

- For discrete random variables, we compute the variance as

$$\mathbb{V}(X) = \sum_{\text{all } x} (x - \mu_X) p(x).$$

- By simple algebra, we can rewrite this expression as

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mu_X^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2.$$

- The standard deviation of X is $sd(X) = \sigma_X = \sqrt{\mathbb{V}(X)}$. The advantage of the standard deviation over the variance is that it is expressed in the same units as the random variable X .
- The coefficient of variation is $CV = \sigma_X / \mu_X$. The CV is not affected by the units in which X is measured.

Variance/Standard Deviation

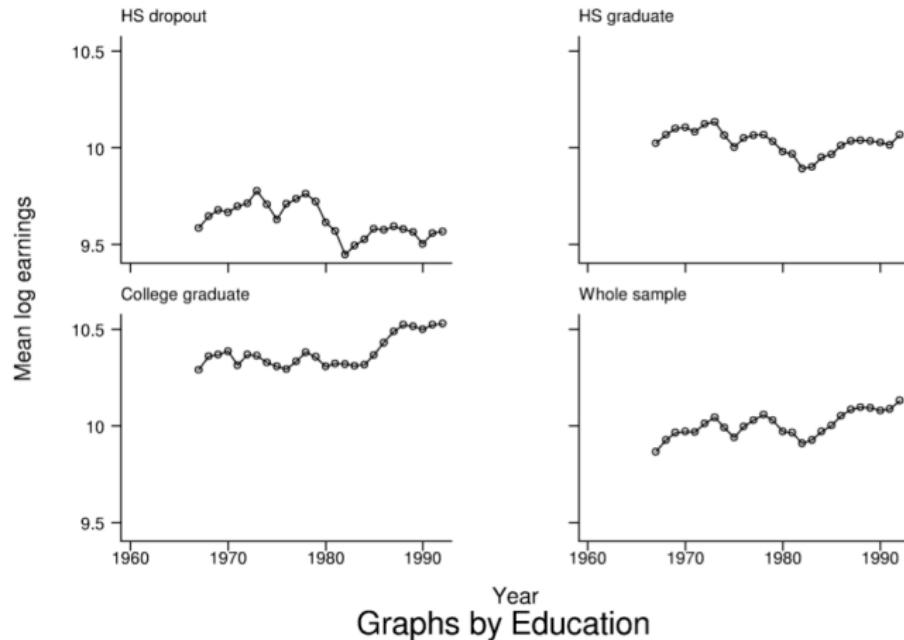
- When do we care about the variance/standard deviation?
- Imagine the following situation:
 - I give you the distribution of temperature in Princeton on February 1st;
 - I ask you to predict which temperature we will have at noon on February 1st;
 - We agree that you will pay to me

$$\$ (\text{actual temperature} - \text{your guess})^2$$

- Example: if you predict $30^{\circ}F$ and the true temperature is $15^{\circ}F$, you will pay to me $\$15^2 = \225
- We saw in the previous class that your best possible guess is the **mean** of the distribution of the temperature in Princeton on February 1st.
- If you choose the mean as your guess, then the variance of the distribution of the temperature in Princeton on February 1st is the expected number of dollars you will have to pay me.

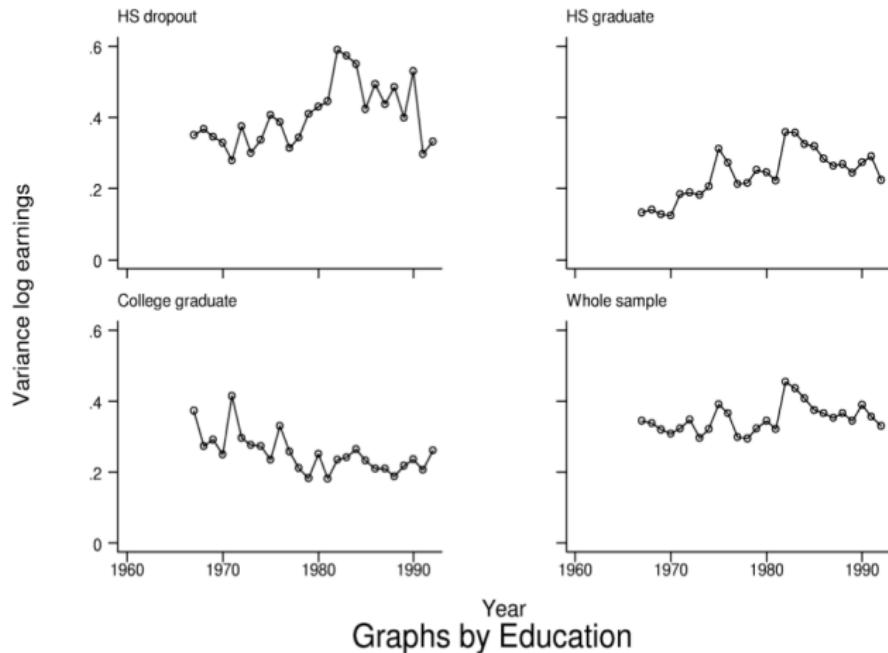
Variance/Standard Deviation

- From Meghir and Pistaferri (2004)



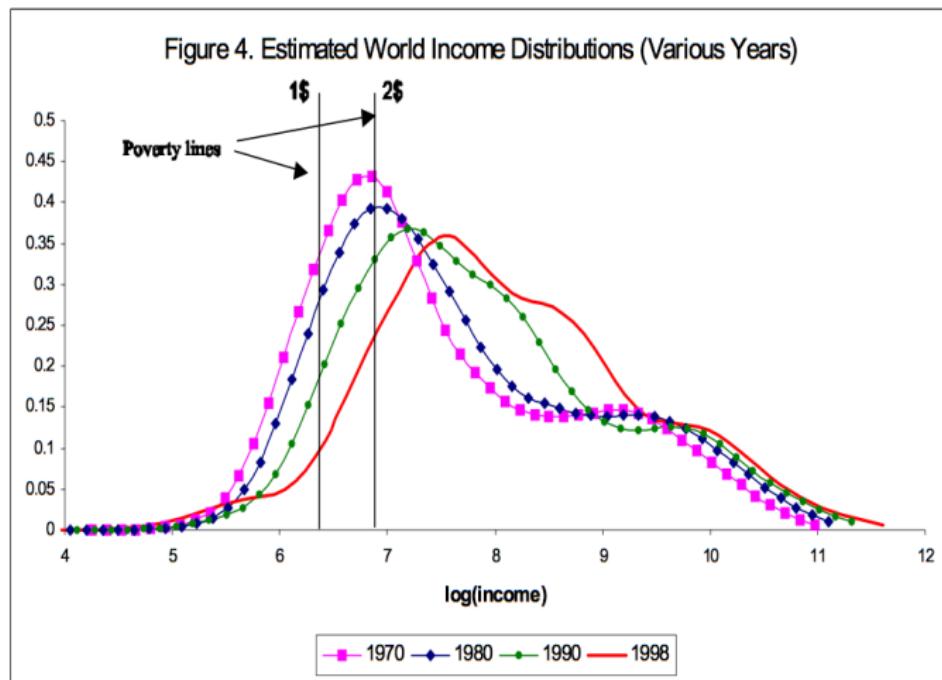
Variance/Standard Deviation

- From Meghir and Pistaferri (2004)



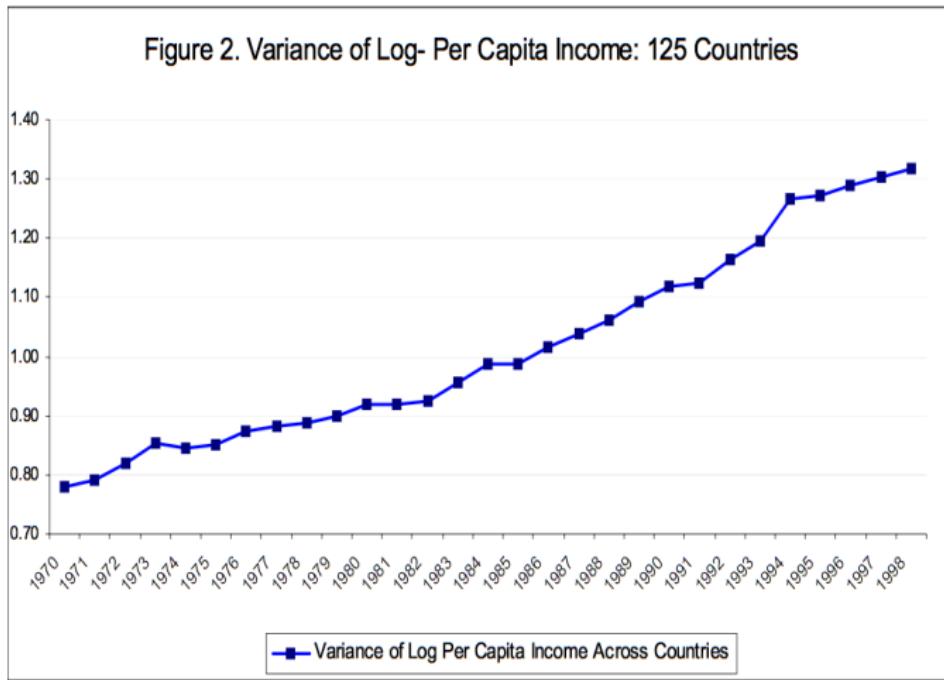
Variance/Standard Deviation

- From Sala-i-Martin (2002)



Variance/Standard Deviation

- From Sala-i-Martin (2002)



Variance/Standard Deviation

- Examples:

- Let $X \sim \text{Bernouilli}(p)$,

$$\begin{aligned}\mathbb{V}(X) &= \mathbb{E}[(X - \mu_X)^2] = \sum_{x=0}^1 (x - p)^2 p(x) = (0 - p)^2(1 - p) + (1 - p)^2 p \\ &= p^2 - p^3 + p + p^3 - 2p^2 = p(1 - p); \\ &= \mathbb{E}(X^2) - \mu_X^2 = \left[\sum_{x=0}^1 x^2 p(x) \right] - p^2 = 0^2(1 - p) + 1^2 p - p^2 = p(1 - p)\end{aligned}$$

- Flip a fair coin two times and denote X the number of heads.

$$\mathbb{V}(X) = \sum_{x=0}^2 (x - \mu_X)^2 p(x) = (0 - 1)^2 \frac{1}{4} + (1 - 1)^2 \frac{1}{2} + (2 - 1)^2 \frac{1}{4} = \frac{1}{2}.$$

- Let $X \sim \text{Uniform}(-1, 3)$,

$$\begin{aligned}\mathbb{V}(X) &= \int_{-1}^3 (x - \mu_X)^2 \frac{1}{4} dx = \int_{-1}^3 (x - 1)^2 \frac{1}{4} dx = \frac{1}{12} \left[(x - 1)^3 \right]_{-1}^3 \\ &= \frac{1}{12} (8 - (-8)) = \frac{16}{12} = \frac{4}{3}\end{aligned}$$

Variance/Standard Deviation

- If $g(\cdot)$ is a function of X , then

$$\mathbb{V}[g(X)] = \int_{-\infty}^{\infty} (g(x) - \mathbb{E}[g(x)])^2 f(x) dx$$

- If a and b are constants and $g(\cdot)$ is a function of X , then

$$\begin{aligned}\mathbb{V}[a + bg(X)] &= \int_{-\infty}^{\infty} (a + bg(x) - a - b\mathbb{E}[g(x)])^2 f(x) dx \\ &= \int_{-\infty}^{\infty} (b(g(x) - \mathbb{E}[g(x)]))^2 f(x) dx \\ &= b^2 \int_{-\infty}^{\infty} (g(x) - \mathbb{E}[g(x)])^2 f(x) dx \\ &= b^2 \mathbb{V}(g(X))\end{aligned}$$

$$sd[a + bg(X)] = b \times sd(g(X)).$$

- If $g(X) = a$, where a is a constant term, then

$$\mathbb{V}(X) = \int_{-\infty}^{\infty} (g(x) - \mathbb{E}[X])^2 f(x) dx = (a - a)^2 = 0.$$

MOMENTS OF DISTRIBUTION OF SCALAR RANDOM VARIABLE: SKEWNESS AND KURTOSIS

Skewness and Kurtosis

- The skewness is defined as:

$$\text{Skewness} = \frac{\mathbb{E}[(X - \mu_X)^3]}{\sigma_X^3}.$$

If skewness > 0 , then the distribution of X has a long/fat right tail.

If skewness < 0 , then the distribution of X has a long/fat left tail.

If skewness $= 0$, then the distribution of X is symmetric around μ_X .

- The kurtosis is defined as:

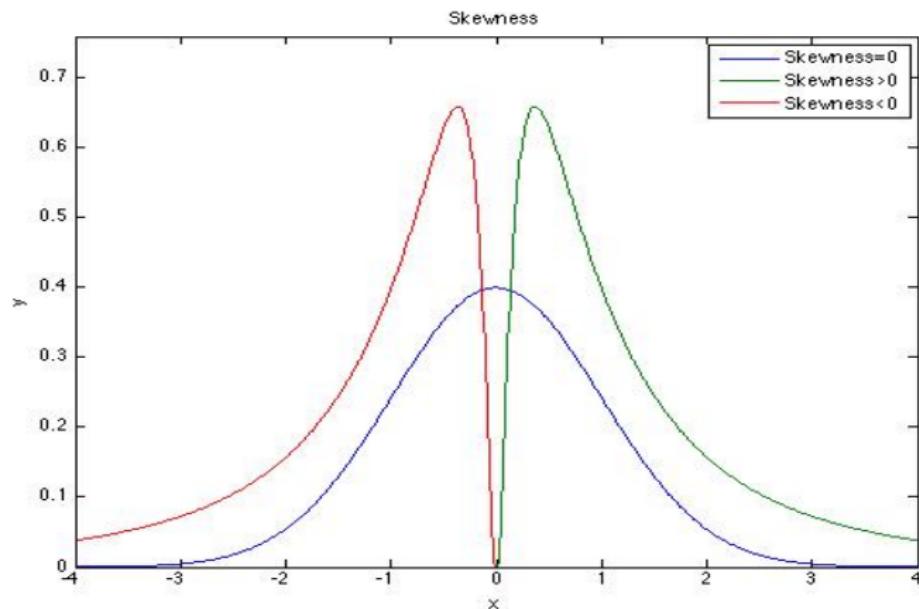
$$\text{Kurtosis} = \frac{\mathbb{E}[(X - \mu_X)^4]}{\sigma_X^4}.$$

If kurtosis > 3 , then the distribution of X has fatter tails than the normal.

If kurtosis < 3 , then the distribution of X has thinner tails than the normal.

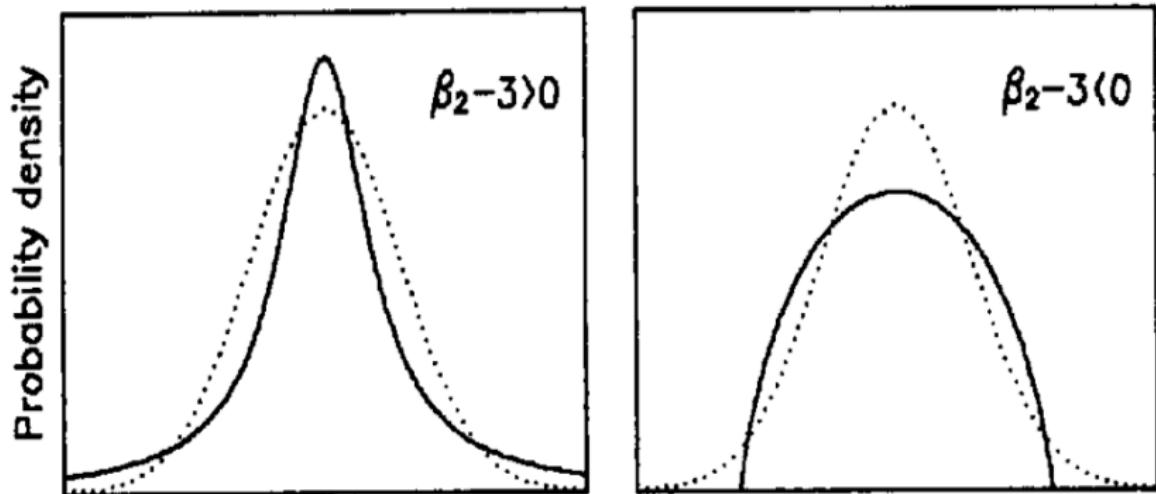
For more info: <http://www.columbia.edu/~ld208/psymeth97.pdf>

Skewness



Kurtosis

- From DeCarlo (1997)



The dotted lines show normal distributions.

Kurtosis

- From DeCarlo (1997)

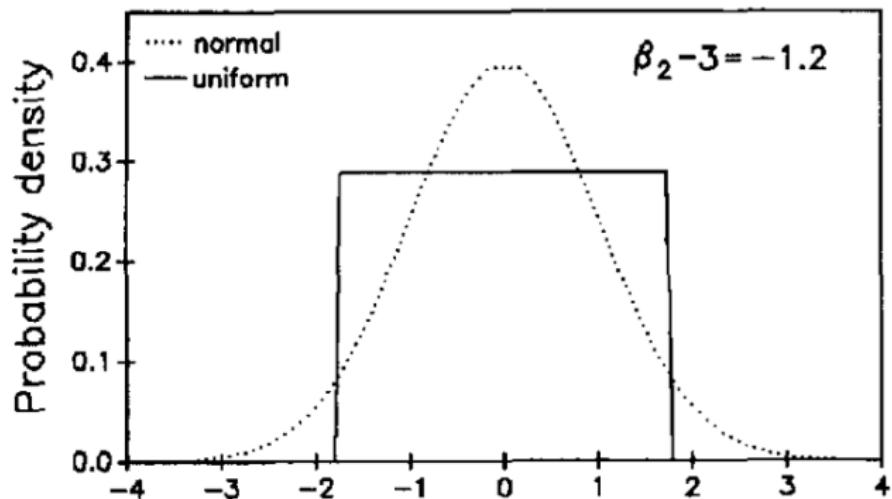


Figure 3. The uniform distribution and the normal distribution, both with a variance of unity.

Exercise

- Compute

- mean
- variance
- skewness
- kurtosis

for a random variable X with the following probability distribution function

$$P(X) = \begin{cases} 0.4 & \text{if } X = 0, \\ 0.6 & \text{if } X = 3. \end{cases}$$

MOMENTS OF JOINT DISTRIBUTION OF TWO RANDOM VARIABLES: COVARIANCE AND CORRELATION COEFFICIENT

Covariance and Correlation Coefficient

- Besides moments that describe the marginal distributions of Y and X , we can also compute measures of how Y and X are related to each other.
- These two measures are:
 - Covariance:

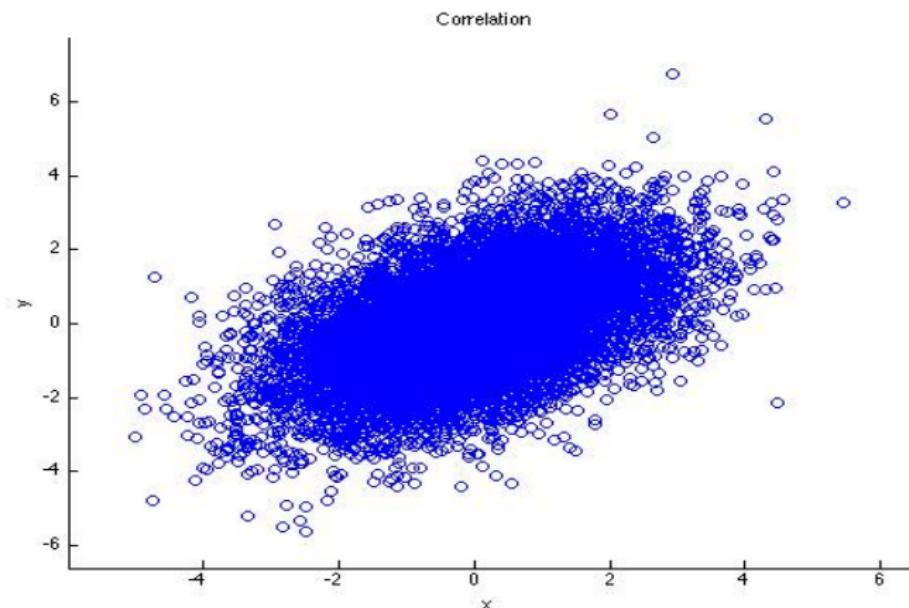
$$\mathbb{C}(X, Y) = \sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

- Correlation coefficient:

$$\rho_{XY} = \frac{\mathbb{C}(X, Y)}{\sqrt{\mathbb{V}(X)}\sqrt{\mathbb{V}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

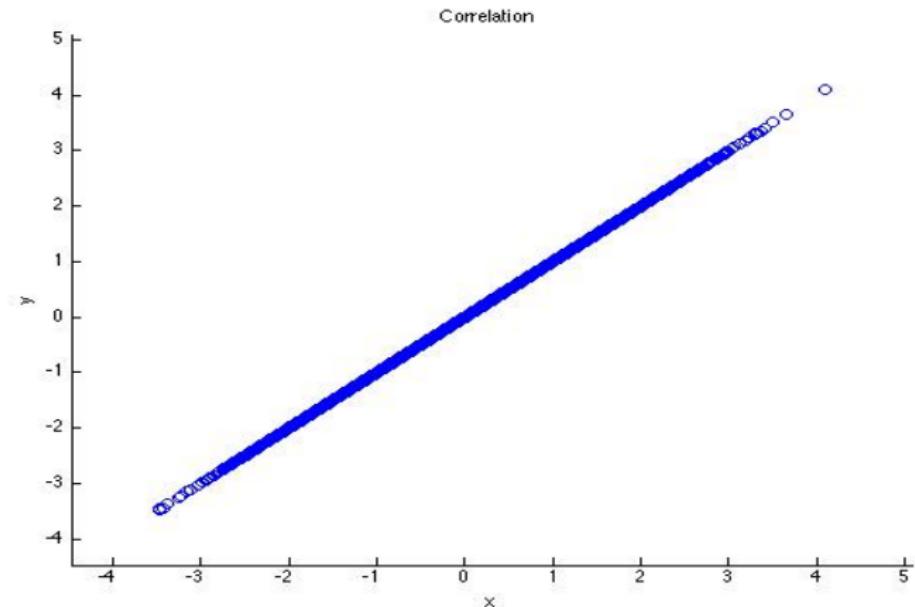
- The main advantage of ρ_{XY} is that it is invariant to the units of X and Y . It satisfies: $-1 \leq \rho_{XY} \leq 1$. Specifically:
 - ① $\rho_{XY} = 1$ if and only if $Y = aX + b$, for some positive a .
 - ② $\rho_{XY} = -1$ if and only if $Y = aX + b$, for some negative a .
 - ③ $\rho_{XY} = 0$ if and only if $Y = aX + b$, for $a = 0$.

Covariance and Correlation Coefficient



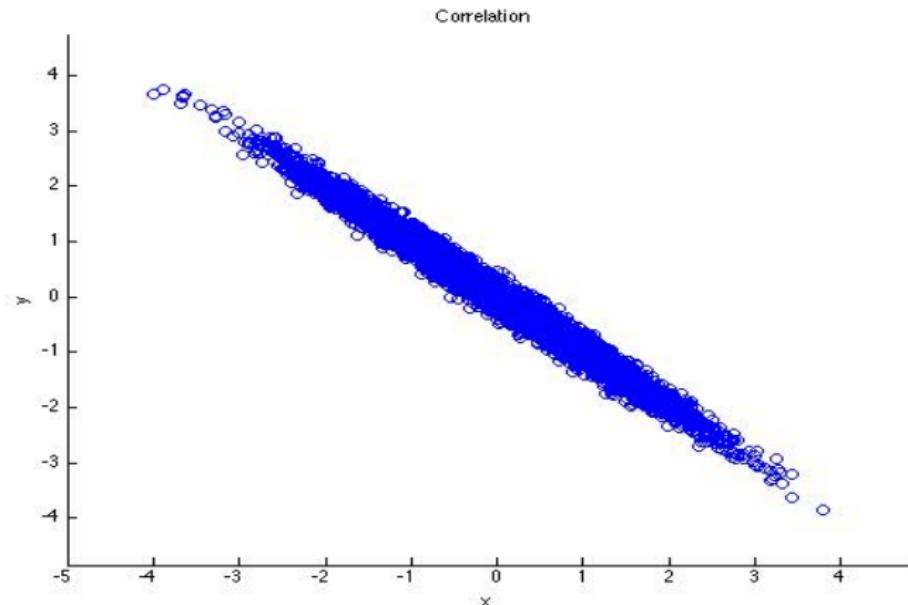
In this figure, $\rho_{XY} = 0.5$.

Covariance and Correlation Coefficient



In this figure, $\rho_{XY} = 1$.

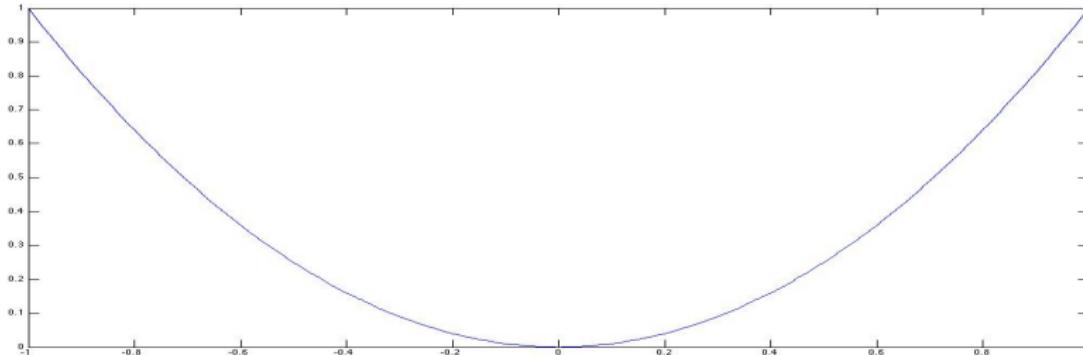
Covariance and Correlation Coefficient



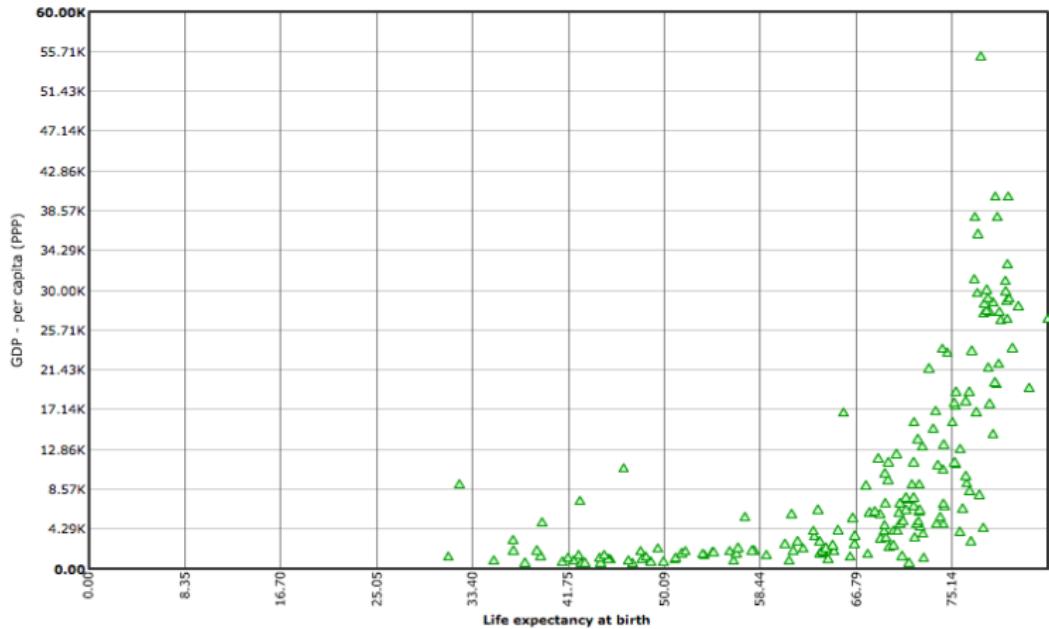
In this figure, $\rho_{XY} = -0.99$.

Covariance and Correlation Coefficient

- Correlation and covariance measure the strength of only the linear relationship between two variables.
- *Correlation does not describe curved relationships between variables, no matter how strong they are*
- Example. Imagine a random variable $X \sim \text{Uniform}[-1, 1]$ and a random variable $Y = X^2$. The correlation coefficient between X and Y is 0. However, a scatterplot of (X, Y) shows a very clear non-linear relationship between both variables:

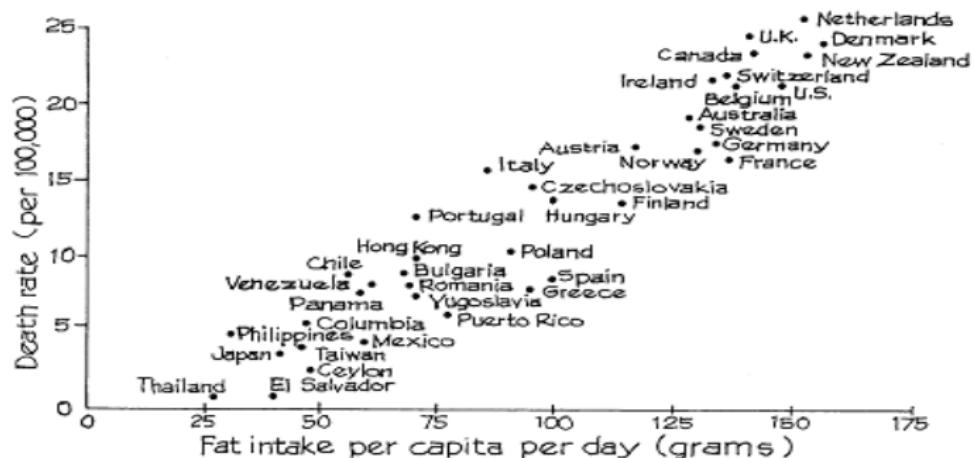


Example of Non-linear Relationship



Example of Non-linear Relationship

Figure 8. Cancer rates plotted against fat in the diet, for a sample of countries.



Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

(Death rate due to breast cancer on the vertical axis)

Other Examples

- Correlation coefficients:
 - between daily maximum temperatures at New York and Boston (in 06/2005)
0.51
 - between income and education for men age 25-64 in the United States (from 2005 CPS)
0.42
 - between rate of cigarette smoking (per capita) and rate of deaths from lung cancer across countries (Richard Doll, 1955)
0.70
 - between height and weight (from Health and Nutrition Examination Survey)
0.40
- Remember that correlation does **not** reflect causation.

MOMENTS OF JOINT DISTRIBUTION OF TWO RANDOM VARIABLES: GENERAL FUNCTIONS OF X AND Y

Moments of Function of Random Variables

- In general

$$\mathbb{E}[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dx dy$$

$$\mathbb{V}[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (h(x, y) - \mathbb{E}[h(X, Y)])^2 f(x, y) dx dy$$

- However, for some particular functions $h(X, Y)$, certain rules simplify the computation of $\mathbb{E}[h(X, Y)]$ and $\mathbb{V}[h(X, Y)]$.

Moments of Function of Random Variables

- If $h(X, Y) = aXY$, then

$$\mathbb{E}[aXY + b] = a\mathbb{E}[XY] + b = a(\mathbb{C}(X, Y) + \mathbb{E}[X]\mathbb{E}[Y]) + b,$$

$$\mathbb{V}[aXY + b] = a^2\mathbb{V}[XY] = a^2\mathbb{E}(XY - \mathbb{E}[XY])^2.$$

- If $h(X, Y) = aXY$ and $\mathbb{C}(X, Y) = 0$, then

$$\mathbb{E}[aXY + b] = a\mathbb{E}[XY] + b = a\mathbb{E}[X]\mathbb{E}[Y] + b,$$

$$\mathbb{V}[aXY + b] = a^2(\mathbb{E}[Y]^2\mathbb{V}(X) + \mathbb{E}[X]^2\mathbb{V}(Y) + \mathbb{V}(Y)\mathbb{V}(X))$$

(if you want to gain practice using these rules, try to prove the last equality)

Moments of Function of Random Variables

- If $h(X, Y) = aX + bY + c$, then

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c,$$

$$\begin{aligned}\mathbb{V}[aX + bY + c] &= \mathbb{V}[aX + bY] \\ &= a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] + 2ab\mathbb{C}(X, Y).\end{aligned}$$

- If $h(X, Y) = aX - bY + c$, then

$$\mathbb{E}[aX - bY + c] = a\mathbb{E}[X] - b\mathbb{E}[Y] + c,$$

$$\begin{aligned}\mathbb{V}[aX - bY + c] &= \mathbb{V}[aX - bY] \\ &= a^2\mathbb{V}[X] + b^2\mathbb{V}[Y] - 2ab\mathbb{C}(X, Y).\end{aligned}$$

- If $h(X, Y) = aX + bY + c$ and $\mathbb{C}(X, Y) = 0$, then

$$\mathbb{V}[aX + bY + c] = \mathbb{V}[aX - bY + c] = a^2\mathbb{V}[X] + b^2\mathbb{V}[Y].$$

MOMENTS OF CONDITIONAL DISTRIBUTION OF A SCALAR RANDOM VARIABLE

Moments of conditional distributions

- We can define the conditional expectation

$$\mathbb{E}[Y|X = x] = \begin{cases} \sum_{\text{all } y} y p(y|X = x) & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} y f(y|X = x) dy & \text{if } Y \text{ is continuous.} \end{cases}$$

and the conditional variance

$$\mathbb{V}[Y|X = x] = \begin{cases} \sum_{\text{all } y} (y - \mathbb{E}[Y|X = x])^2 p(y|X = x) & \text{if } Y \text{ is discrete,} \\ \int_{-\infty}^{\infty} (y - \mathbb{E}[Y|X = x])^2 f(y|X = x) dy & \text{if } Y \text{ is continuous.} \end{cases}$$

- Note that both $\mathbb{E}[Y|X]$ and $\mathbb{V}[Y|X]$ are functions of X (not of Y).
- Useful rules:
 - $\mathbb{E}[g(X)|X] = g(X);$
 - $\mathbb{V}[g(X)|X] = 0$
- More useful rules:
 - $\mathbb{E}[g(X)h(Y)|X] = g(X)\mathbb{E}[h(Y)|X];$
 - $\mathbb{V}[g(X)h(Y)|X] = (g(X))^2\mathbb{V}[h(Y)|X]$

Moments of conditional distributions

- Example for discrete random variable:

	$p_{Y X}(y_j X = 8)$	$p_{Y X}(y_j X = 12)$	$p_{Y X}(y_j X = 16)$
$Y = 25$	$0.15/0.2 = 0.75$	$0.15/0.45 = 0.33$	$0/0.35 = 0$
$Y = 45$	$0.05/0.2 = 0.25$	$0.15/0.45 = 0.33$	$0.10/0.35 = 0.29$
$Y = 70$	$0/0.2 = 0$	$0.15/0.45 = 0.33$	$0.25/0.35 = 0.71$

$$\mathbb{E}[Y|X = 8] = (25 \times 0.75) + (45 \times 0.25) + (70 \times 0) = 30,$$

$$\mathbb{E}[Y|X = 12] = (25 \times 0.33) + (45 \times 0.33) + (70 \times 0.33) = 46.67,$$

$$\mathbb{E}[Y|X = 16] = (25 \times 0) + (45 \times 0.29) + (70 \times 0.71) = 62.75.$$

or, as an alternative example,

$$\mathbb{E}[Y + X|X = 8] = (33 \times 0.75) + (53 \times 0.25) + (78 \times 0) = 38,$$

$$\mathbb{E}[Y + X|X = 12] = (37 \times 0.33) + (57 \times 0.33) + (82 \times 0.33) = 58.67,$$

$$\mathbb{E}[Y + X|X = 16] = (41 \times 0) + (61 \times 0.29) + (86 \times 0.71) = 78.75.$$

LAW OF ITERATED EXPECTATIONS

Law of Iterated Expectations

- The LIE establishes a relationship between marginal and conditional expectations.

$$\mathbb{E}_Y[Y] = \mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]] \text{ and } \mathbb{E}_X[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y]].$$

More generally: $\mathbb{E}_{XY}[h(X, Y)] = \mathbb{E}_X[\mathbb{E}[h(X, Y)|X]]$

- Proof:

$$\begin{aligned}\mathbb{E}_{XY}[h(X, Y)] &= \int \int h(x, y) f_{XY}(x, y) dx dy \\ &= \int \int h(x, y) f_{Y|X}(y|x) f_X(x) dx dy \\ &= \int \int h(x, y) f_{Y|X}(y|x) f_X(x) dy dx \\ &= \int \left[\int h(x, y) f_{Y|X}(y|x) dy \right] f_X(x) dx \\ &= \int [\mathbb{E}_{Y|X}[h(X, Y)|X]] f_X(x) dx = \mathbb{E}_X[\mathbb{E}_{Y|X}[h(X, Y)|X]].\end{aligned}$$

Law of Iterated Expectations

- Example 1. Suppose we draw X from a Bernouilli distribution with success probability p . If $X = 1$, then we draw Y from a uniform distribution on $[-2, 0]$. If $X = 0$, then we draw Y from a uniform distribution on $[2, 6]$. Compute the marginal expectation of Y : $\mathbb{E}_Y[Y]$.

$$\mathbb{E}_Y[Y] = \mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]]$$

where

$$\mathbb{E}_{Y|X}[Y|X] = \begin{cases} \int_{-2}^0 y f_Y(y) dy = \int_{-2}^0 y \frac{1}{2} dy = -1 & \text{if } X = 1, \\ \int_2^6 y f_Y(y) dy = \int_2^6 y \frac{1}{4} dy = 4 & \text{if } X = 0, \end{cases}$$

and

$$\begin{aligned} \mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]] &= \sum_{x=0}^1 \left\{ \mathbb{E}_{Y|X}[Y|X=x] p(x) \right\} \\ &= (-1 \times p) + (4 \times (1 - p)) = 4 - 5p. \end{aligned}$$

Law of Iterated Expectations

- Example 2. Suppose we draw $X \sim \text{Uniform}(0, 1)$. After we observe $X = x$, we draw $Y|X = x \sim \text{Uniform}(x, 1)$. Compute $\mathbb{E}_Y[Y]$.

$$\mathbb{E}_Y[Y] = \mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]]$$

where

$$\begin{aligned}\mathbb{E}_{Y|X}[Y|X = x] &= \int_x^1 y f_{Y|X}(y|X = x) dy = \int_x^1 y \frac{1}{1-x} dy = \frac{1}{1-x} \left[\frac{y^2}{2} \right]_x^1 \\ &= \frac{1-x^2}{2(1-x)} = \frac{1+x}{2}\end{aligned}$$

and

$$\mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]] = \int_0^1 \frac{1+x}{2} dx = \frac{1}{2} + \frac{1}{2} \int_0^1 x dx = \frac{1}{2} + \frac{1}{2} \times \frac{1}{2} = \frac{3}{4}.$$

WWS 507c: Quantitative Analysis

Lecture 8: Families of Continuous Distributions

Princeton University

October 9, 2014

REVIEW OF LECTURE 7

Review of Lecture 7

- Write as a function of $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{V}(X)$, $\mathbb{V}(Y)$, and $\mathbb{C}(X, Y)$:

$$\mathbb{E}(3 + 4X) =$$

$$\mathbb{V}(3 + 4X) =$$

$$\mathbb{E}(X^2) =$$

$$\mathbb{E}(XY) =$$

$$\mathbb{V}(X + Y) =$$

$$\mathbb{V}(X - Y + 50) =$$

$$\mathbb{V}(3X - Y + 50) =$$

INTRODUCTION

Introduction

- In previous lectures, we saw that **any** function $f(x)$ may be a PDF as long as $f(x) \geq 0$, for all $-\infty < x < \infty$, and $\int_{-\infty}^{\infty} f(x) = 1$.
- Many of these functions $f(x)$ may be classified into what are called **families** of probability density functions.
- For e.g., the **family** of uniform distributions includes **all** the PDFs of the kind

$$\text{Uniform}(a, b) = f(x; (a, b)) = \begin{cases} 0 & \text{if } x < a, \\ \frac{1}{b-a} & \text{if } a \leq x < b, \\ 0 & \text{if } b \leq x, \end{cases}$$

for any possible values of a and b as long as $b > a$.

- We say that the uniform distribution depends on two **parameters**: a and b .
 - the lower limit of its support: a
 - the upper limit of its support: b
- E.g. the distribution $\text{Uniform}(0, 1)$ is one particular member of this family.

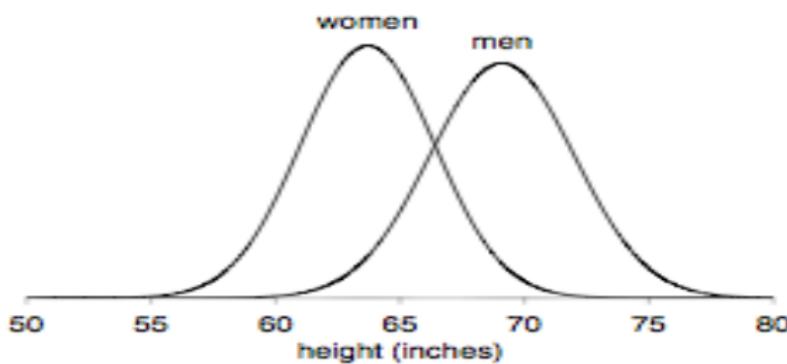
Introduction

- The particular **families** of density functions we will focus on in this lecture are:
 - Normal,
 - Chi-squared,
 - F,
 - Student's t,
- During the rest of the semester, we will encounter many variables whose probability distribution belongs to one of these families.
- Therefore, we will spend some time today learning the properties of the random variables whose probability distribution belongs to one of these families.

UNIVARIATE NORMAL DISTRIBUTION: INTRODUCTION

Normal Distribution

- When is a random variable X normally distributed? When can we say that a random variable X is normal?
- Informal (and imprecise) answer: whenever its probability density function looks like a bell. E.g. distribution of heights of men and women in the U.S:



Women. mean: 63.7; s.d.: 2.7

Men. mean: 69.1; s.d.: 2.9

Normal Distribution

- Formal (and precise –but possibly less useful–) answer: whenever its PDF is equal to

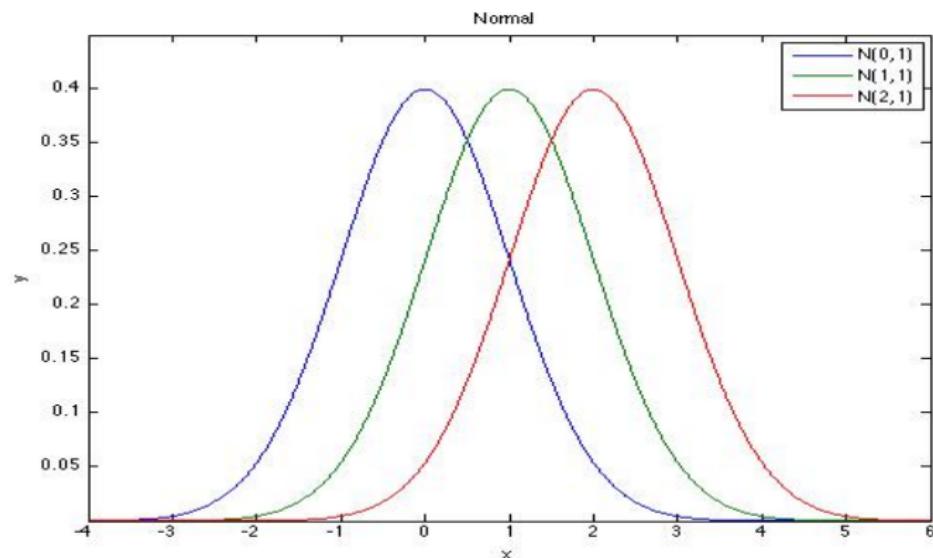
$$f(x; (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty \leq x \leq \infty.$$

for some particular values of the **parameters** μ and σ .

- What are these μ and σ that operate as parameters of the normal family?
 - the parameter μ is precisely the mean of X
 - the parameter σ is precisely the standard deviation of X
- If a random variable X is distributed normally with mean μ and standard deviation σ , then we write $X \sim \mathbb{N}(\mu, \sigma^2)$.
- Remember, if two normal random variables X_1 and X_2 have different values of μ or σ , then their PDFs will be different (even if they all look bell-shaped).

Normal Distribution

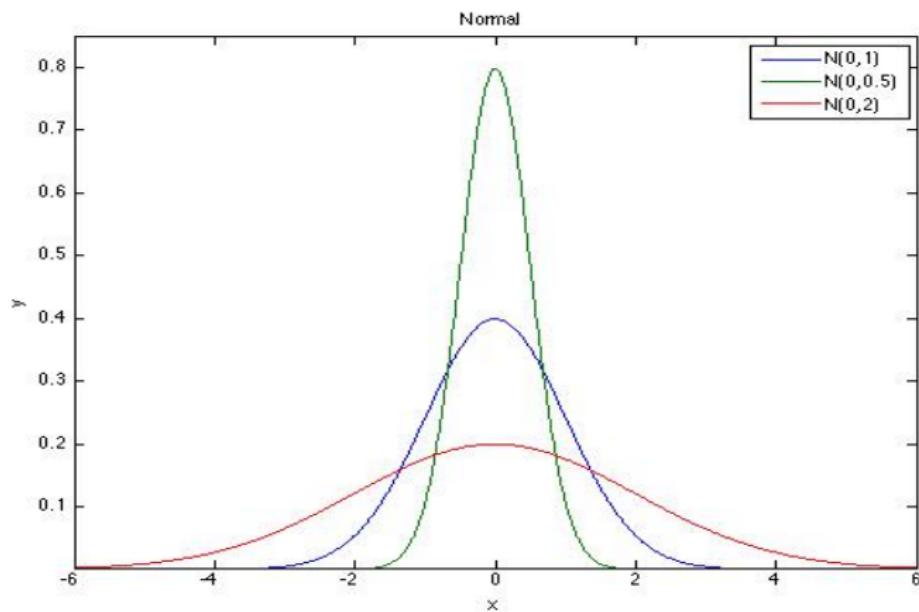
- How does the PDF of a normal random variable change as we modify μ ?



Note that the PDF is symmetric and centered around μ . As we increase μ , we shift the distribution to the right.

Normal Distribution

- How does the PDF of a normal random variable change as we modify σ ?



As we increase σ , the distribution becomes more spread.

UNIVARIATE NORMAL DISTRIBUTION: HISTORY

Short History of the Normal Distribution

- A fun summary of the history of the normal distribution may be found at:
https://www.maa.org/sites/default/files/pdf/upload_library/22/Allendoerfer/stahl96.pdf
- I will provide here a quick summary.
- We saw in previous lectures that the binomial distribution tells us the probability that we observe x 1s in a sequence of n independent Bernoulli trials

$$P(X = m|n, p) = \frac{n!}{n!(n-m)!} p^m (1-p)^{n-m}$$

- This is something that Pascal and Fermat had figured out in 1654 after a few years of fruitful correspondence between the two of them.
- However, one limitation with this formula was that, in a world without computers, it was very hard to compute it when n was a large number.

Short History of the Normal Distribution

- For example, in 1712, Gravesande wanted to compute the probability of the observed number of male and female births in London during the period 1629-1710, given the assumption that the probability of a male birth was 0.5.
- In his data, there were approximately 11,000 births per year with approximately 6,000 male births. Therefore, using the binomial formula, one needed to perform some very onerous tasks as computing

$$\frac{11,000!}{6,000! 5,000!}$$

or computing

$$0.5^{6,000}$$

- Given the impossibility of computing these numbers, a mathematician, De Moivre, started to search for an approximation in 1721.

Short History of the Normal Distribution

- In 1735, De Moivre found that the formula

$$f(x; (\mu, \sigma)) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

yields a very good approximation to the probability of observing m success in n Bernoulli trials, as long as n was a large number and

$$x = \frac{m}{n} - p \quad \mu = pn \quad \sigma = np(1-p)$$

Short History of the Normal Distribution

- This figure both the actual distribution and the normal approximation to the number of 1s in 50 Bernoulli trials with $p = 0.3$.

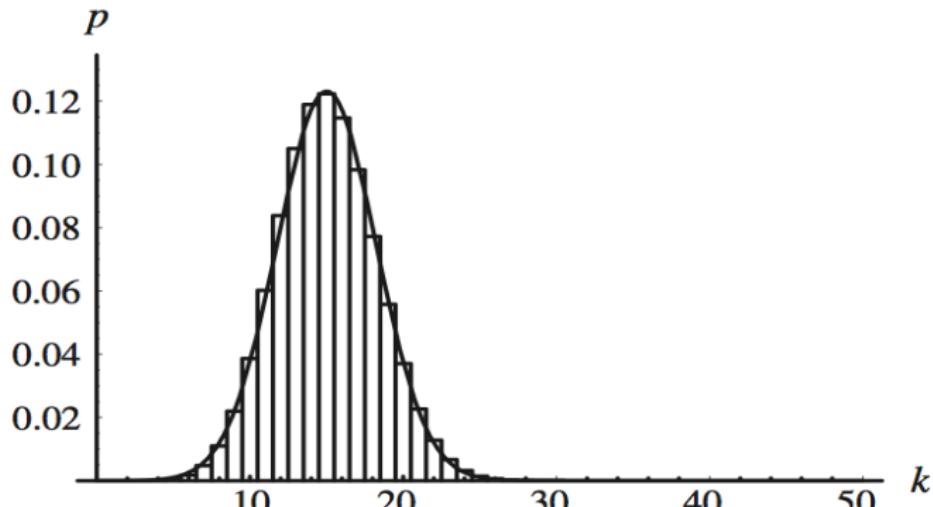


Figure 2 An approximation of binomial probabilities

Short History of the Normal Distribution

- Since its discovery by De Moivre, this formula started to appear everywhere.
- First, it appeared in astronomy.
- In 1600, Pascal noticed that different measurements of the same quantity (e.g. right ascension to Mars) were yielding widely different numbers.
- This was the first realization of the problem of measurement error (it will reappear later in the semester...).
- Faced with multiple measurements of the same quantity, astronomers had to find one representative measurement.
- Then it started a median versus average controversy that lasted for many centuries.
- In the 19th century, a young mathematician, Gauss, argued that the best summary of measures contaminated by error was the average and showed that the distribution of the errors around this average followed exactly the normal distribution!

Short History of the Normal Distribution

- This figure shows the distribution of the deviations of different measurements with respect to its mean.

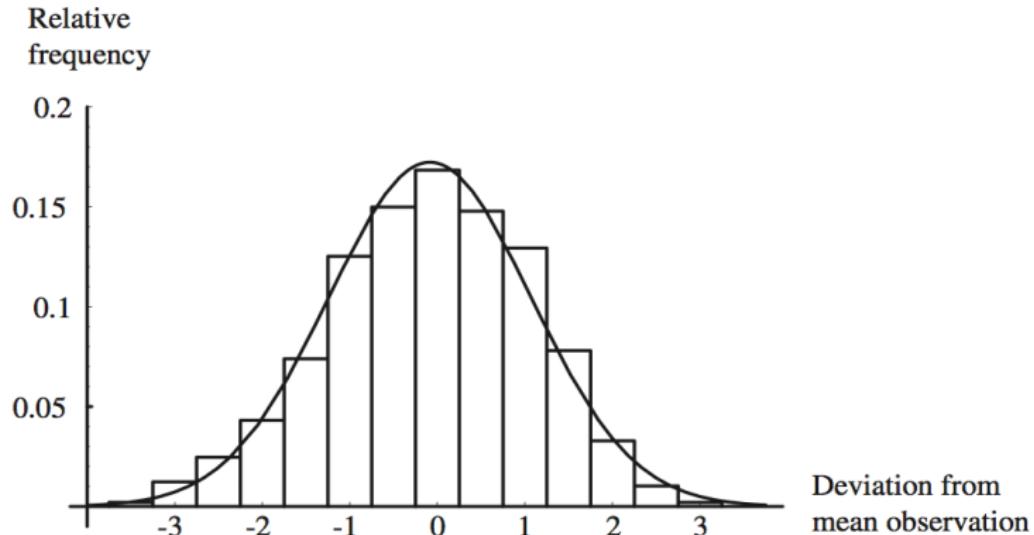


Figure 7 Normally distributed measurements

Short History of the Normal Distribution

- And the first person to apply the normal distribution to the social sciences was Adolphe Quetelet in 1846.
- He began his career as an astronomer but, at some point in his life, he started to collect large quantities of data on crime, divorce rate, human height, etc.
- As an example, he measured the chests of 5,738 Scottish soldiers and he found that the measures also followed a normal distribution!

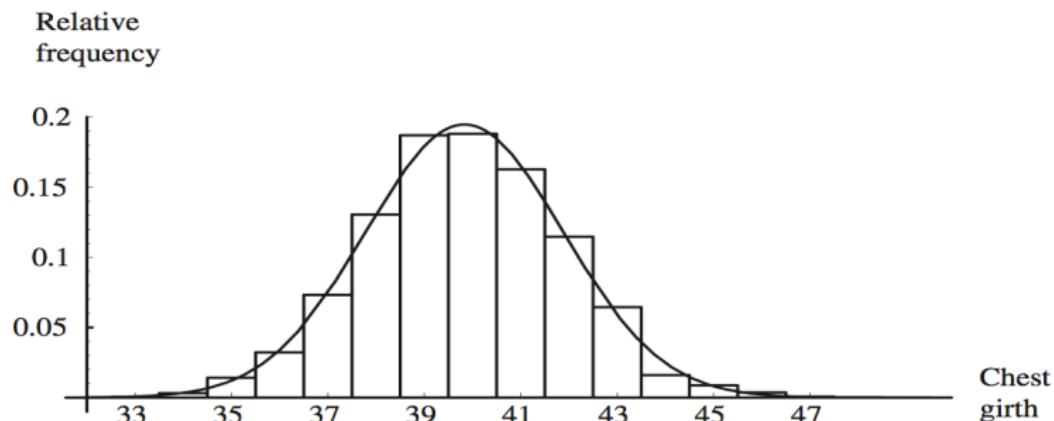


Figure 8 Is this data normally distributed?

Short History of the Normal Distribution

- In case you are curious about the exact chest measurements....

TABLE 1: Chest measurements of
Scottish soldiers

Girth	Frequency
33	3
34	18
35	81
36	185
37	420
38	749
39	1,073
40	1,079
41	934
42	658
43	370
44	92
45	50
46	21
47	4
48	1
	5,738

Short History of the Normal Distribution

- And it was not only the chest measurements, but also heights and weights seemed to follow a normal distribution...
- And, why the name *normal*?
- For a long while, it was customary to refer to this distribution as *Gaussian curve* (in honor of Gauss).
- However, there was some dispute about whether Gauss or another statistician, Laplace, had been the first to discover such curve.
- So a third statistician, Karl Pearson, unaware that the true discovery of this curve was made by De Moivre, wrote in 1920:

“Many years ago [in 1983], I called the Laplace-Gaussian curve the normal curve, which name, it avoids the international question of priority,”

UNIVARIATE NORMAL DISTRIBUTION: PROPERTIES

Normal Distribution

- All the normal random variables (independently of their values of μ and σ) have the following property:
 - ① the area under the normal curve between $-\sigma$ and σ is about 0.68.
 - ② the area under the normal curve between -2σ and 2σ is about 0.95.
 - ③ the area under the normal curve between -3σ and 3σ is about 0.997.
- Remember that this is totally equivalent to saying:

$$P(\mu - \sigma \leq Z \leq \mu + \sigma) = 0.68,$$

$$P(\mu - 2\sigma \leq Z \leq \mu + 2\sigma) = 0.95,$$

$$P(\mu - 3\sigma \leq Z \leq \mu + 3\sigma) = 0.997.$$

Properties of Univariate Normal Random Variables

Any linear transformation of a normal random variable is also normal.

- Assume a and b are constants/real numbers, then

If $\left\{ \begin{array}{l} (1) \text{ } X \text{ is normally distributed} \\ (2) \text{ } Y = a + bX \end{array} \right\}$ then Y is normally distributed.

- Also, using general rules about expectations and variances of functions of random variables (Lecture 7), we know that

$$\mathbb{E}[Y] = \mathbb{E}[a + bX] = a + b\mathbb{E}[X],$$

$$\mathbb{V}[Y] = \mathbb{V}[a + bX] = b^2\mathbb{V}[X],$$

- Therefore, if X is normally distributed with mean μ_X and standard deviation σ_X and $Y = a + bX$, then Y is normally distributed with mean $a + b\mu_X$ and standard deviation $b\sigma_X$:

$$\text{if } X \sim \mathbb{N}(\mu_X, \sigma_X^2), \text{ then } Y \sim \mathbb{N}(a + b\mu_X, b^2\sigma_X^2) = \mathbb{N}(\mu_Y, \sigma_Y^2).$$

Standard Normal Distribution

- Of all the possible normal random variables, there is one that is special!
- It is so special that it has its own name: *standard* normal random variable.
- The standard normal random variable is the normal random variable that has mean equal to 0 and standard deviation equal to 1:
 - $\mu = 0$
 - $\sigma = 1$
- It is standard to reserve the letter Z to denote this random variable

$$Z \sim \mathbb{N}(0, 1).$$

- The PDF of the standard normal random variable is

$$f(Z; (0, 1)) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}Z^2\right), \quad -\infty \leq Z \leq \infty.$$

Standard Normal Distribution

- Let's focus for now on the standard normal distribution.
- For some numbers b , how would you compute $P(-\infty \leq Z \leq b)$?
- If we were to use the methods that we have been learning in previous lectures, we would do the following:

$$P(-\infty \leq Z \leq b) = \int_{-\infty}^b f(Z; (0, 1)) dZ = \int_{-\infty}^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}Z^2\right) dZ.$$

- But this integral has no (analytic) solution!
- Luckily, one can use a computer to compute this integral for many many different values of b . The results from this calculation are published in what is usually called *the statistical table for the standard normal distribution*.
- To get practice, use the statistical table uploaded in blackboard to compute
 - $P(-\infty \leq Z \leq -1)$, if $Z \sim \mathbb{N}(0, 1)$.
 - $P(-2 \leq Z \leq 0.5)$, if $Z \sim \mathbb{N}(0, 1)$.

Normal Distribution

- For normal random variables X that are **not** the standard normal random variable, how can we compute the probability that X is between some number a and some other number b ?
- Example, if $X \sim \mathbb{N}(\mu, \sigma^2)$ and either $\mu \neq 0$ or $\sigma \neq 1$, how can you compute

$$P(a \leq X \leq b) ?$$

- We do this in two steps:
 - First, we use algebra to transform this expression into an expression that depends on the standard normal random variable Z

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right).$$

- Second, we use *the statistical table for the standard normal distribution*.
- To get practice, let's compute
 - $P(-25 \leq Z \leq 0)$, if $X \sim \mathbb{N}(-10, 5)$.
 - $P(-25 \leq Z \leq 0)$, if $X \sim \mathbb{N}(5, 10)$.

Exercise

- The US population aged between 25 and 54 is approximately 100 M. Half of them are women. The distribution of heights for men in the US is $N(69.1, 2.9)$, and the one for women is $N(63.7, 2.7)$. Assume that only 2% of men and 1% would accept a job offer from the U.S. Fire Administration. Besides, the U.S. Fire Administration hiring regulations establishes a minimum height of 67 inches to work for the U.S. Fire department. Answer the following questions:
 - ➊ How large is the pool of workers between 25 and 54 years old from which the U.S. Fire Administration might be able to hire? Note: this is the pool of workers that would accept a job offer from the U.S. Fire Administration **and** are above 67 inches tall.
 - ➋ Indicate any additional assumption you had to impose to answer the previous question.

Exercise

- Define

- $N = \text{US population aged between 25 and 54.}$

$$N = 100 \text{ M}$$

- $N_H = \text{US population between 25 and 54 years old that might be hired by the U.S. Fire Administration.}$
 - This is the number we want to compute.
 - $H = \text{dummy variable equal 1 if individual might be hired by the US Fire Administration.}$
 - $A = \text{dummy variable equal 1 if individual would accept a job offer from U.S. Fired Administration.}$
 - $F = \text{dummy variable for female.}$

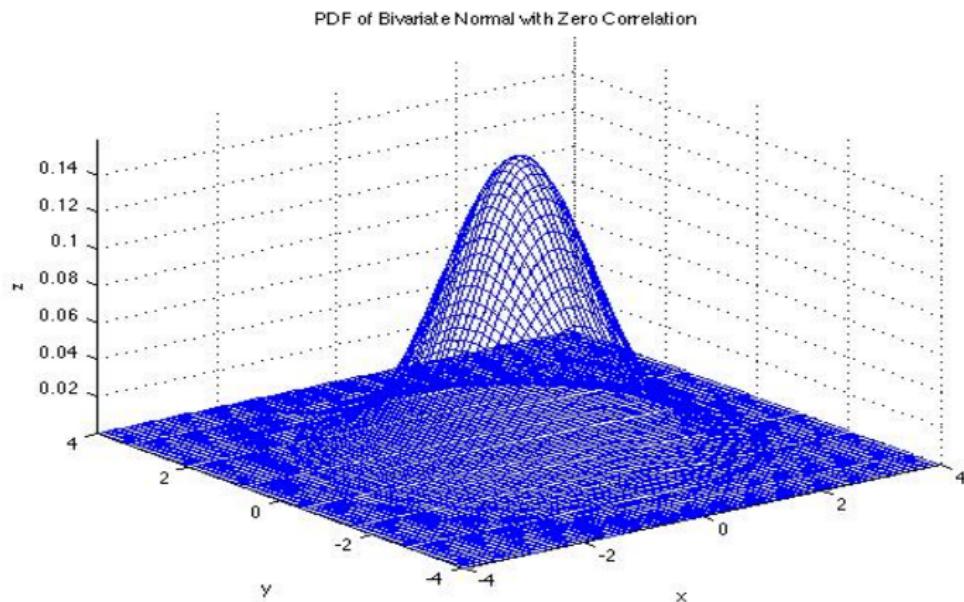
Exercise

$$\begin{aligned}N_H &= N \times \mathbb{P}[H = 1] = N \times \mathbb{E}[\mathbb{1}\{H = 1\}] = N \times \mathbb{E}_{F,A}[\mathbb{E}[\mathbb{1}\{H = 1\}|F, A]] = \\N \left(\mathbb{P}(F = 0, A = 0) \mathbb{E}[\mathbb{1}\{H = 1\}|F = 0, A = 0] + \mathbb{P}(F = 1, A = 0) \mathbb{E}[\mathbb{1}\{H = 1\}|F = 1, A = 0] + \right. \\&\quad \left. \mathbb{P}(F = 0, A = 1) \mathbb{E}[\mathbb{1}\{H = 1\}|F = 0, A = 1] + \mathbb{P}(F = 1, A = 1) \mathbb{E}[\mathbb{1}\{H = 1\}|F = 1, A = 1] \right) = \\N \left(\mathbb{P}(F = 0, A = 0) \times 0 + \mathbb{P}(F = 1, A = 0) \times 0 + \right. \\&\quad \left. \mathbb{P}(A = 1|F = 0) \mathbb{P}(F = 0) \mathbb{P}[h > 67|F = 0] + \mathbb{P}(A = 1|F = 1) \mathbb{P}(F = 1) \mathbb{P}[h > 67|F = 1] \right) = \\100 \left(0.02 \times 0.5 \times P(Z > \frac{67 - 69.1}{2.9}) + 0.01 \times 0.5 \times P(Z > \frac{67 - 63.7}{2.7}) \right) = \\1 \times P(Z > -0.72) + 0.5 \times P(Z > 1.22) = \\&\quad \Phi(0.72) + 0.5 \times (1 - \Phi(1.22)) = \\0.7642 + 0.5 \times (1 - 0.8888) = \\0.8198 \text{ millions of potential applicants} = \\819,800 \text{ potential applicants}\end{aligned}$$

BIVARIATE NORMAL DISTRIBUTION

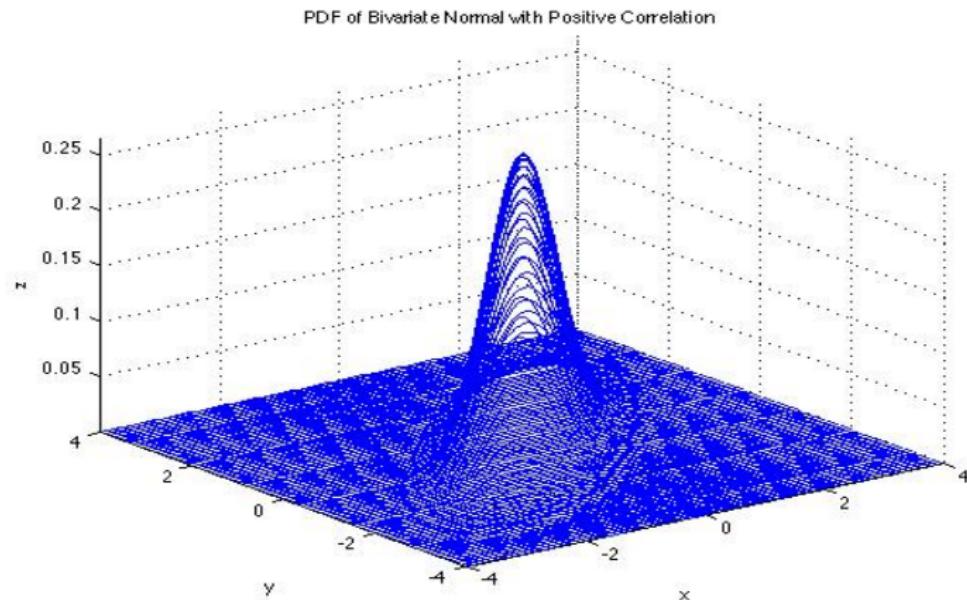
Bivariate Normal Distribution

- How does the bivariate normal distribution looks like?
- If both random variables are independent, it looks like:



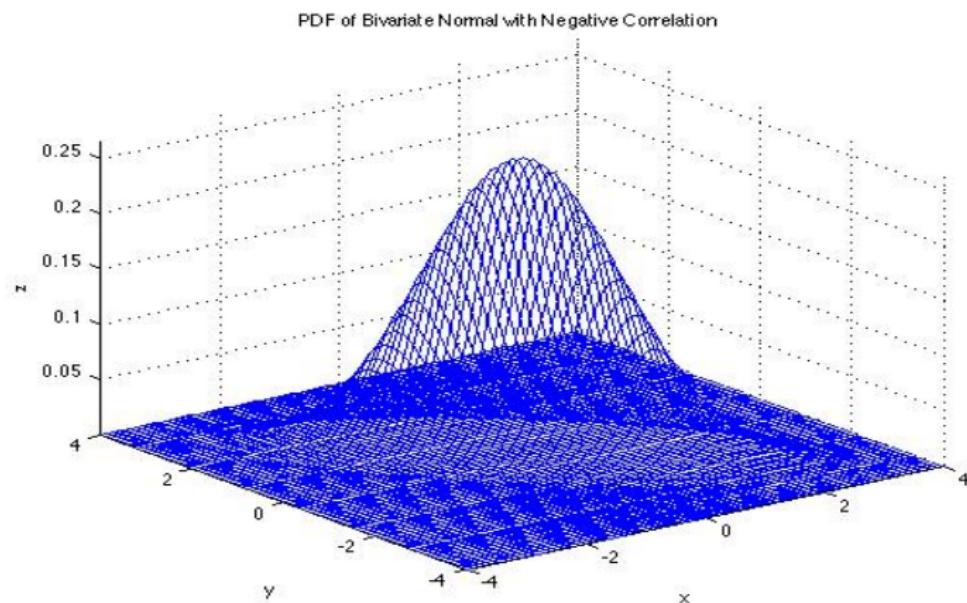
Bivariate Normal Distribution

- If both random variables are positively correlated, it looks like:



Bivariate Normal Distribution

- If both random variables are negatively correlated, it looks like:



Properties of Bivariate Normal Distribution

- Property 1 of the bivariate normal distribution:
If (X, Y) are jointly normal, then both X and Y are scalar normal random variables
- Does it work in the opposite direction? This means, if X and Y are scalar normal random variables, is it true that (X, Y) are jointly normal?
- If X and Y are scalar normal random variables and $\rho_{XY} \neq 0$, then it is not generally true that (X, Y) are jointly normal.
- If X and Y are scalar normal random variables and $\rho_{XY} = 0$, then (X, Y) are jointly normal.

Properties of Bivariate Normal Distribution

- Property 2 of the bivariate normal distribution:

If (X, Y) are jointly normal, then both the conditional distribution $X|Y = y$ and the conditional distribution $Y|X = x$ are normal, for any value of y and x .

In particular,

$$X|(Y = y) \sim \mathbb{N}(\mu_{X|Y}, \sigma_{X|Y}^2) = \mathbb{N}\left(\mu_X + \frac{\sigma_{XY}}{\sigma_Y^2}(y - \mu_Y), (1 - \rho_{XY}^2)\sigma_X^2\right),$$

- Notice that:

- the mean of the distribution $X|(Y = y)$ depends **linearly** on y .
- the variance of the distribution $X|(Y = y)$ does not depend on y .
- the variance of the distribution decreases in the correlation coefficient of X and Y .

Exercise

- PubCorr is a Spanish NGO that has been keeping track of all corruption scandals in Spain during the last 25 years. Using data from all court sentences covering cases of misappropriation of public funds by publicly elected officials, it was has discovered the following joint distribution:

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \left[\begin{pmatrix} 30 \\ 5 \end{pmatrix}, \begin{pmatrix} 25 & 10 \\ 10 & 4 \end{pmatrix} \right]$$

where X denotes millions of euros that a civil servant manages and Y denotes millions of euros that he misappropriates.

The data for this exercise have no bearing with reality...of course...

Exercise

- Compute the correlation coefficient between X and Y .
- Which distribution is more concentrated around its mean?
- What is the expected quantity misappropriated by an elected official that manages 50 million euros?
- What is the probability that an elected official that manages 50 million euros misappropriates between 7 and 8 million euros?
- Imagine that the covariance between X and Y was equal to 5 (instead of 10), compute again the probability that an elected official that manages 50 million euros misappropriates between 7 and 8 million euros?

Properties of Bivariate Normal Distribution

- Property 3 of the bivariate normal distribution.

X and Y are independent random variables if and only if $\rho_{XY} = 0$.

- Remember that, in general, for random variables that are not normally distributed
 - if two random variables are independent, then $\rho_{XY} = 0$,
 - but, on the contrary, ρ_{XY} being equal to 0 does not imply that X and Y are independent.
- The special characteristic of normal random variables is that, if $\rho_{XY} = 0$, then we can conclude that X and Y are independent.

Properties of Bivariate Normal Distribution

- Property 4 of bivariate normal distribution.

If (X, Y) are jointly normal, then any linear combination of X and Y is also normally distributed

- Corollary: any linear combination of two independent normal random variables is also normal; i.e.

If $\left\{ \begin{array}{l} (1) X \sim \mathbb{N}(\mu_X, \sigma_X) \\ (2) Y \sim \mathbb{N}(\mu_Y, \sigma_Y) \\ (3) \rho_{XY} = 0 \\ (4) W = aX + bY \end{array} \right\}$ then W is normally distributed,

and

$$\mathbb{E}[aX + bY] = \mu_W = a\mu_X + b\mu_Y$$

$$\begin{aligned} \mathbb{V}[aX + bY] &= \sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2. \end{aligned}$$

A VERY important result

- This result is an implication of Property 4.
- Suppose Z_i , for $i = 1, \dots, n$, are independent standard normal random variables. Let

$$W = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i.$$

Then W is also a standard normal random variable.

- As an exercise, try to prove this result by yourself.
- Something to remark is that **the distribution of W does NOT depend on n** . The random variable W follows the same distribution independently of whether $n = 1$ or n is a very large number.

CHI-SQUARED DISTRIBUTION

Chi-Squared Distribution

- The different distributions that belong to the chi-squared family differ only in one parameter. This parameter is called **degrees of freedom** and it is usually denoted as n . We denote a Chi-Squared distribution with n degrees of freedom as

$$\chi_n^2$$

- The distribution of a Chi-Squared distribution with n degrees of freedom is identical to the distribution of the sum of n independent squared standard normal random variables:

$$W = \sum_{i=1}^n Z_i^2,$$

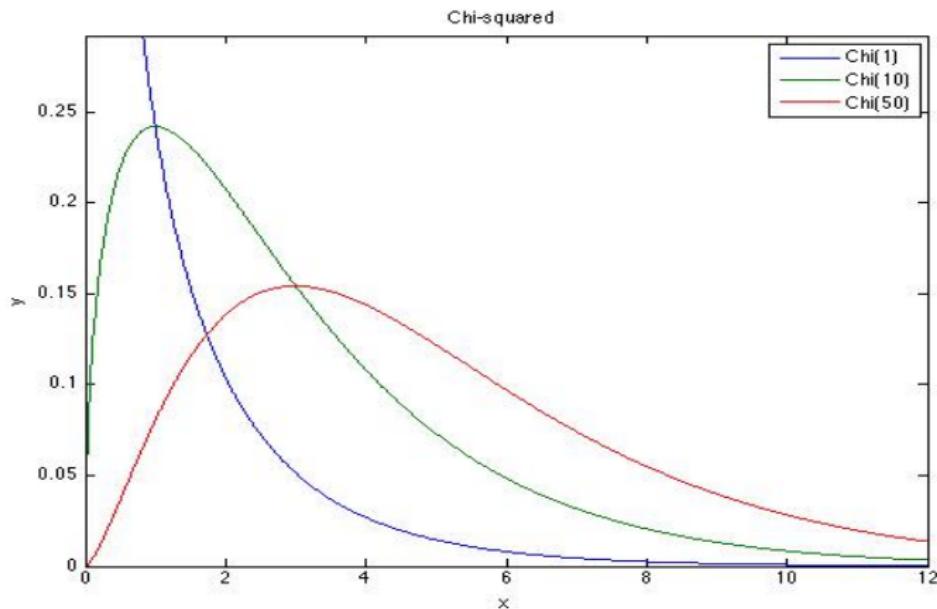
where Z_i , for $i = 1, \dots, n$, are independent standard normal random variables.

- Therefore

$$\mathbb{E}[\chi_n^2] = \mathbb{E}\left[\sum_{i=1}^n Z_i^2\right] = \sum_{i=1}^n \mathbb{E}[Z_i^2] = \sum_{i=1}^n \mathbb{V}[Z_i] = \sum_{i=1}^n 1 = n.$$

Chi-Squared Distribution

- How does the Chi-Squared distribution look like?



Chi-Squared Distribution

- If W is distributed Chi-Squared with n degrees of freedom (i.e. $W \sim \chi_n^2$), and both a and b are constants, how do we compute $P(a \leq W \leq b)$?
- We use a table that, for a very large set of possible degrees of freedom n and numbers b , indicates the value of

$$P(b \leq W) = 1 - F(b).$$

- Note that the table for the standard normal random variable gives the CDF. Conversely, the table for the Chi-Squared gives you $1 - CDF$.
- Exercise. Use the table that contains the CDF for the chi-squared distribution and compute:
 - ① $P(W \geq 20.6)$, if $W \sim \chi_{30}^2$.
 - ② $P(1.6 \leq W \leq 12.8)$, if $W \sim \chi_5^2$.

F DISTRIBUTION

F Distribution

- The different distributions that belong to the F family differ in two parameters. These two parameters are called: **degrees of freedom of the numerator** (n_1), and **degrees of freedom of the denominator** (n_2). Therefore, we denote an F distribution as

$$F_{n_1, n_2}$$

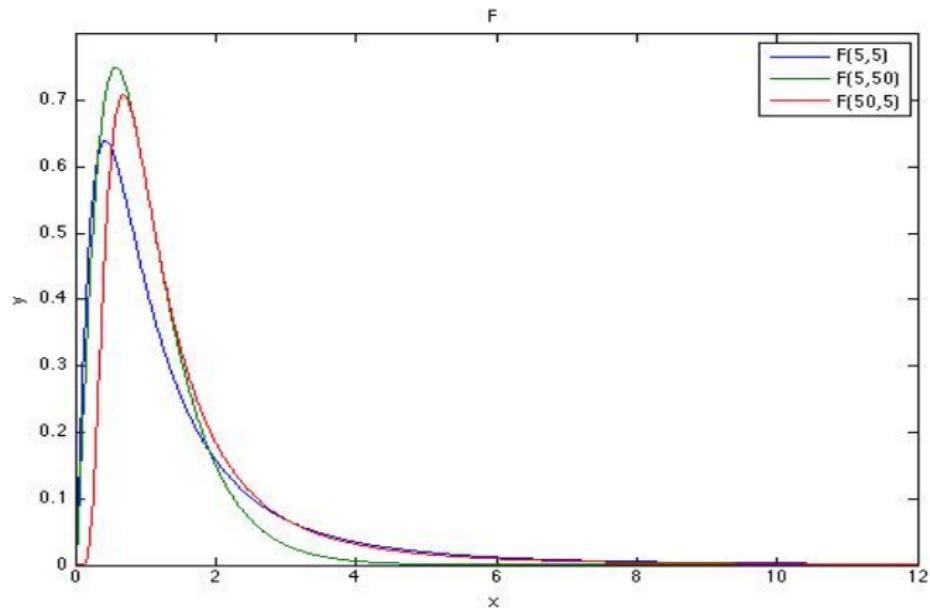
- The distribution of a random variable F that has an F distribution is identical to the distribution of the ratio of two Chi-Squared random variables, divided by their respective degrees of freedom.
- Formally, a random variable F has an F distribution with (n_1, n_2) degrees of freedom if and only if

$$F = \frac{W_1/n_1}{W_2/n_2},$$

where both W_1 and W_2 are independent Chi-Squared random variables with respective degrees of freedom n_1 and n_2 .

F Distribution

- How does the F distribution look like?



STUDENT'S t DISTRIBUTION

Student's t Distribution

- The different distributions that belong to the student's t family differ in one parameter. This parameter is called again **degrees of freedom (n)**. Therefore, we denote a student's t distribution as

$$t_n$$

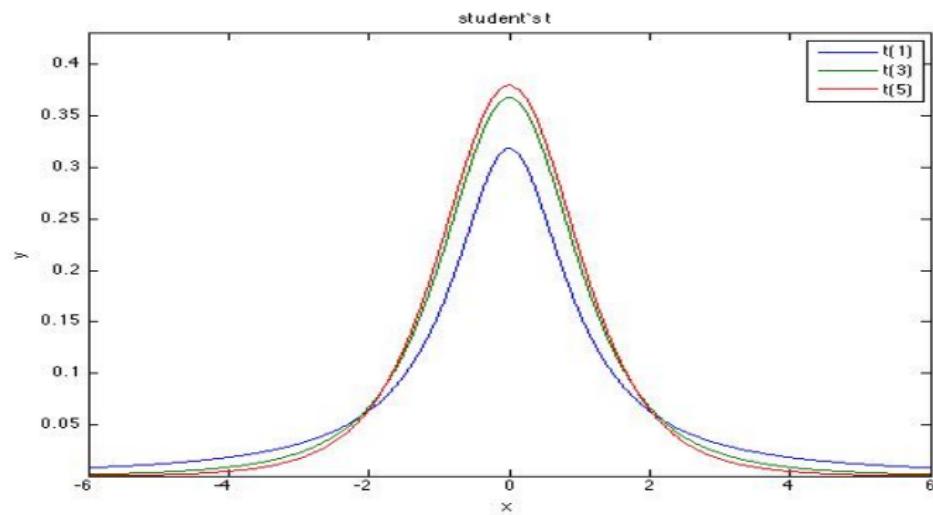
- A random variable t has a Student's t distribution (or, simply, t distribution) with n degrees of freedom if and only if it can be represented as

$$t = \frac{Z}{\sqrt{W/n}},$$

where both Z and W are independent random variables, Z has a standard normal distribution, and W has a χ_n^2 distribution.

Student's t Distribution

- How does the Student's t distribution look like?



For small number of degrees of freedom n , the t distribution is similar to the standard normal but with fatter tails. As the degrees of freedom get larger and larger, the Student's t distributions gets closer and closer to the standard normal.

WWS 507c: Quantitative Analysis

Lecture 9: Random Sampling and Distribution of the Sample Average

Princeton University

October 16, 2014

RANDOM SAMPLE

Population of Interest and Sample

- I have given you a survey with three questions:
 - How many fiction books have you read during 2014? (random variable X)
 - How many foreign countries did you visit during 2013 and 2014? (random variable Y)
 - How many concerts did you attend during 2014? (random variable W)
- Each of these three questions defines a random variable. Random variables capture information about each of the individuals included in the **population of interest**. In the case of the survey above, the population of interest is all of you: the students taking 507c.
- A **sample** is a subset of the individuals included in the population of interest.
- Examples of possible samples on this population are:
 - subset of the class whose family name starts with a T;
 - subset of the class who was born abroad;
 - subset of the class who plays an instrument;

Population of Interest and Sample

- The population of interest and sample may be different in different settings.
- Setting 1. Gallup Daily Tracking conducts 1,500 interviews per day among landline and cell phones across the U.S. for its political and economic survey:

[http://www.gallup.com/poll/113980/
Gallup-Daily-Obama-Job-Approval.aspx](http://www.gallup.com/poll/113980/Gallup-Daily-Obama-Job-Approval.aspx)

The population of interest is the set of all adults living in the U.S. However, even though Gallup wants to know which fraction of all adults living in the U.S. approve or disapprove of the job Barack Obama is doing as president, it does not ask each of these adults every day. Gallup only asks 1,500 adults that it selects at random. These 1,500 adults form a sample.

- The random variable of interest is: dummy variable indicating approval for Obama's job.

Population of Interest and Sample

- Setting 2. The National Cancer Institute performs clinical trials to test different cancer treatments.
- The population of interest is the set of individuals that:
 - have a certain type or stage of cancer;
 - have received a certain kind of therapy in the past;
 - are in a certain age group.
- The sample is the set of individuals who, belonging to the population of interest, voluntarily join the clinical trials.
- The size of the sample depends on the clinical trial phase:

<http://www.cancer.gov/clinicaltrials/learningabout/Taking-Part-in-Cancer-Treatment-Research-Studies/page3>
- The random variable of interest is: dummy variable indicating recovery from cancer.

Random Sample: Definition

- A random sample is a sample that results from a procedure that extracts individuals from the population of interest:
 - one by one;and such that,
 - in each extraction, every individual in the population of interest has the same probability of being extracted.
- Imagine that you have one ball for each individual in a population of interest and that you introduce all the balls in one box. You would obtain a random sample of size, for example, 10, if you extract 10 balls **with replacement** and, in each draw, **all balls have equal probability of being extracted**.
- Using a die, we can extract a random sample of size 5 from the set of students in 507c. Recording the value of X , Y , and Z for each of the 5 students of the sample, we obtain:

$$\{(X_1, Y_1, W_1), (X_2, Y_2, W_2), \dots, (X_5, Y_5, W_5)\}.$$

Random Sample: Properties

- **Property 1.** The numeric value attached to each draw or element of a random sample is a random variable.
- Example: the number of fiction books that the first element of the random sample has read, X_1 , is a random variable.
 - it is a variable because it is a number
 - it is random because, if you were to repeat the random phenomenon that has extracted that individual from the sample (e.g. extract another ball from the box; roll again the dice), then you may obtain a different number.
- Property 1 is not exclusive to random samples. It will also hold for alternative sampling procedures.
 - Even if we extract balls without replacement, each element of the resulting sample is still a number and *ex ante* uncertain. Therefore, each element of the sample constructed in this way is still a random variable.

Random Sample: Properties

- **Property 2.** The random vectors

$$(X_1, Y_1, W_1), (X_2, Y_2, W_2), \dots, (X_5, Y_5, W_5).$$

are **independent of each other** and **identically distributed**.

- We usually write this property as: $(X_1, Y_1, W_1) \dots, (X_n, Y_n, W_n)$ are *iid*, n being the sample size.
- Importantly, this implies that X_1, \dots, X_5 are *iid*, Y_1, \dots, Y_5 are *iid*, and W_1, \dots, W_5 are *iid*. Each of the variables that we collect for one observation will be independent and identically distributed of the corresponding variable for each of the other observations in the sample.
- Example: the number of foreign countries visited by the second element of the random sample, Y_2 , will be independent from and identically distributed to the number of foreign countries visited by the fourth element of the random sample, Y_4 .

Random Sample: Properties

- Importantly, property 2 does **not** imply that, for example, X_1 and Y_1 , or X_1 and W_1 are *iid*.
- Example: property 2 does not indicate that the distribution of the number of fiction books read in 2014 is independent of the number of concerts attended during 2014.
- If we had drawn the balls from the box without replacement, then the different elements of the random sample would not be identically distributed and would not be independent. Therefore, random sampling requires sampling with replacement. Random sampling requires that each individual in the population has the possibility of being sampled more than once.
- However, when populations are very large, sampling without replacement generates probability distributions for each random draw that are very very similar to those that would be generated sampling with replacement.

Consumer Price Index (CPI)

- In order to compute the CPI, the Bureau of Labor Statistics (BLS) collects a sample of prices for different goods. For example, one of the goods considered in the CPI is *hot chocolate drink*. For our purposes, let's focus on this good.
- In order to compute the CPI for *hot chocolate drink*, the BLS takes every month a sample of prices.
- The population of interest is the thousands of retail stores or service establishments that sell *hot chocolate drink* for a price in the U.S.
- The sample the BLS uses to compute the CPI is a subset of these stores.
- On each store, the BLS defines random variables that capture prices. One of these random variables is the price of the hot chocolate drink. Let's denote this variable as p .
- Which is the marginal distribution of p_1 ? For each possible price, its probability is given by the frequency with which that price is observed across all the stores selling hot chocolate.

Consumer Price Index (CPI)

- Which is the distribution of p_2 ?
- If the prices are sampled with replacement, it will be the same as the distribution of p_1 : for each price, its probability is given by the frequency with which that price is observed across all stores selling hot chocolate.
- However, the BLS does not sample prices with replacement!
- Sampling without replacement, the distribution of p_2 will be: for each possible price, its probability is given by the frequency with which that price is observed across all stores selling hot chocolate, *excluding the store that was sampled in the first place*.
- Note that, if the set of stores selling hot chocolate is large, then sampling with replacement is very similar to sampling without replacement.
- More generally, if the sample size is small relative to the size of the population, then sampling with and without replacement yields very similar distributions for the random variable capturing each element of the sample.

Consumer Price Index (CPI)

- Independently of how the sample is obtained, once the BLS has a set of prices for hot chocolate drink

$$\{p_1, p_2, \dots, p_n\},$$

where n is the set of sampled establishments serving hot chocolate drink, what does the BLS do with these prices?

- It computes its average. The average of a random variable across the elements included in a sample is called the **sample average**.
- In the remaining of this lecture, we will study the **properties of the sample average of a random sample**.

Why the Emphasis on Random Samples?

- Hypothesis 1: “A random sample will be representative of the population”.
- Is hypothesis 1 correct?

NO, NO, NO, NO, NO, NO, NO, NO, NO, NO

- Due to chance, the distribution of a random variable in a random sample of finite size n might turn out to be very very different from its distribution in the population from which the sample was drawn.
- For example, it is possible that if you toss a fair die n times, all tosses will come up six. In this case, the distribution of a random variable Y denoting the outcome of the die is very different in the sample ($p(Y = 6) = 1$) and in the population ($p(Y = y) = 1/6$ for $y = 1, 2, 3, 4, 5, 6$).

Why the Emphasis on Random Samples?

- The correct answer is that we put so much emphasis on random samples because “only” when the sample is random we know how to use the information captured in the sample to learn something about the population.¹
- Example. “Only” if CBS News asked a random sample of voters, we can learn something about the entire population of voters from the following numbers

Most Important Issue in Your Vote

Among registered voters	All	Reps	Dems	Inds
Economy	34%	31%	37%	35%
Health care	17%	9%	23%	17%
Terrorism	16%	19%	15%	14%
Immigration	13%	18%	9%	12%
Budget deficit	9%	10%	9%	9%
International conflicts	7%	7%	5%	9%

SOURCE: CBS News

¹The “only” is a bit strong, but it is true that only advanced statistical methods will deal with samples other than random samples.

Why the Emphasis on Random Samples?

- The set of techniques that allow us to use the information from a random sample to learn something about the population from which the sample was drawn is called **inference**.
- We will start to learn about inference techniques in the next lecture.
- We will rely heavily on things we have learned in previous lectures.
- One last comment...You might wonder how Gallup or CBS News obtain a random sample of 1,000 individuals from a population of hundreds of millions of individuals.
- No, you are right, they do not have a box with hundred of millions of balls from which they take 1,000 of them. They have computers that generate random numbers between 1 and millions very quickly. For example, check the following website

<http://www.random.org/>

SAMPLE AVERAGE: FINITE SAMPLE PROPERTIES

Sample Average: Definition

- For the rest of this lecture, assume that we observe a random sample of size n , $Y_i, i = 1, \dots, n$, with $\mathbb{E}[Y_i] = \mu_Y$ and $\mathbb{V}[Y_i] = \sigma_Y^2$.
- We define the sample average or sample mean as the average of the random variables $Y_i, i = 1, \dots, n$:

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

- The finite sample properties are properties of the sample average that will hold independently of the size of the sample (i.e. no matter how small or large the sample is).

Sample Average: Properties

- Property 1 of the sample average

The sample average is a random variable

- This implies that, if we take different random samples of size n from the same population and compute the sample average, \bar{Y} , for each of these samples,

$$\{\bar{Y}_s, s = 1, \dots, S\},$$

we will very likely obtain a different value \bar{Y}_s for each sample s .

- If the BLS were to simultaneously take two random samples of 1000 observed prices from the distribution of prices of hot chocolate drink for all US establishments, the average of each sample will likely be different.
- This implies that, as every other random variable, the sample mean computed over a random sample has a PDF, a CDF, a mean, a variance, etc.
- The rest of this lecture focuses on describing the properties of the distribution of the sample mean of a random sample.

Sample Average: Properties

- Property 2 of the sample average

The expectation of the sample mean of a random sample of $F(Y)$ is the same as the expectation of Y :

$$\mathbb{E}[\bar{Y}] = \mu_Y.$$

- Proof. \bar{Y} is a linear combination of Y_1, \dots, Y_n ; therefore, using the rules of expectations of functions of random variables, we know

$$\mathbb{E}[\bar{Y}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i] = \frac{1}{n} \sum_{i=1}^n \mu_Y = \frac{1}{n} n \mu_Y = \mu_Y$$

- Note on terminology. Instead of saying “we draw a random sample from a population for which a random variable Y has a CDF $F(Y)$ ”, it is standard to say “we draw a random sample of $F(Y)$ ”.
- Note that the expectation of the sample average of a random sample does not depend on sample size.

Sample Average: Properties

- Property 3 of the sample average

The variance of the sample mean of a random sample of $F(Y)$ is the same as the variance of Y divided by the sample size, n :

$$\mathbb{V}[\bar{Y}] = \frac{\sigma_Y^2}{n}$$

- Proof. \bar{Y} is a linear combination of Y_1, \dots, Y_n ; therefore, using the rules over expectations of functions of random variables, we know

$$\begin{aligned}\mathbb{V}[\bar{Y}] &= \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n^2} \mathbb{V}\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}[Y_i] = \frac{1}{n^2} \sum_{i=1}^n \sigma_Y^2 = \frac{1}{n^2} n \sigma_Y^2 \\ &= \frac{\sigma_Y^2}{n}.\end{aligned}$$

- Note that the variance of the sample mean decreases with n and goes to 0 as $n \rightarrow \infty$.

Sample Average

- Property 4 of the sample average

The sampling distribution of the sample mean of a random sample of a normal distribution $F(Y)$ is also normal

- If $Y_i, i = 1, \dots, n$ are independent and identically normally distributed, then $\bar{Y} = (1/n) \sum_{i=1}^n Y_i$ is also normally distributed.
- Proof. The distribution of a linear function of independent normal random variables is also normal (see Lecture 8).
- If the Y_i 's are *i.i.d* draws taken from a distribution that is not normal, then \bar{Y} will not be normally distributed. The distribution of \bar{Y} will depend on the particular distribution of the Y_i 's.
- However, if the Y_i 's are *i.i.d* draws taken from a distribution that is not normal and the size of the sample is “large” (i.e. n is “large”), then the distribution of \bar{Y} will be approximately normal (we will come back to this in a few slides).

Sample Average

- Properties 1 to 4 have in common that they apply to any random sample, independently of its size n .
- Putting together properties 1 to 4, we can conclude that:

If \bar{Y} is the sample average of a random sample, $\{Y_i, i = 1, \dots, n\}$, taken from a normal distribution, then

$$\bar{Y} \sim \mathbb{N}(\mu_Y, \frac{\sigma_Y^2}{n}),$$

where $Y_i \sim \mathbb{N}(\mu_Y, \sigma_Y^2)$.

- In the following slides, we will study additional properties of the sample average of a random sample that only apply when the sample size is “large”.

Implications of Properties 1 to 4

- The distribution of the combined Critical Reading and Mathematics SAT scores is approximately normal with mean 1010 and standard deviation 225.
- Properties 1 to 4 imply:
 - First, if we want to predict the average grade for a random sample of students, our best guess is 1010.
 - Second, this guess might be very wrong (sample average might be very different from population mean).
 - Third, the distance between our guess and the actual sample average is likely to be smaller the larger n is.
 - Fourth, if we took multiple samples of size n , the distribution of the different sample means will look normal or bell-shaped around the mean of 1010. This is true independently of how large n is.

SAMPLE AVERAGE: LARGE SAMPLE APPROXIMATIONS

Large-Sample Approximations

- While the properties 1 to 4 are exact, the properties that only apply when the sample size is “large” will be approximations.
- We study here two large-sample approximations:
 - Law of Large Numbers (LLN)
 - Central Limit Theorem (CLT)

Law of Large Numbers

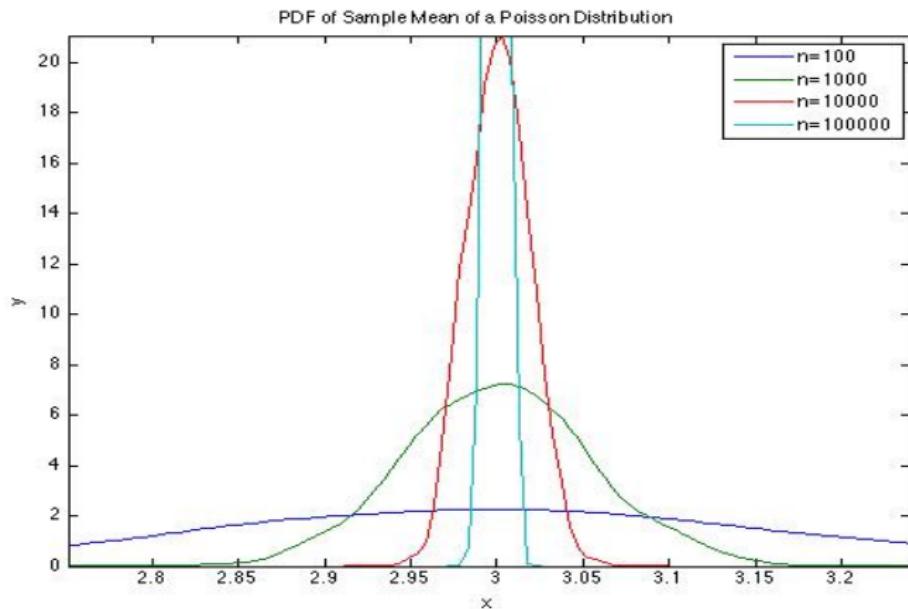
- Draw independent observations at random from any population with finite mean μ . As the number of observations drawn increases, the probability that the mean \bar{Y} differs from the mean μ of the population in more than any fixed number a decreases.
- For any real number a ,

$$P(|\bar{Y}_n - \mu| > a)$$

decreases as the sample size n increases.

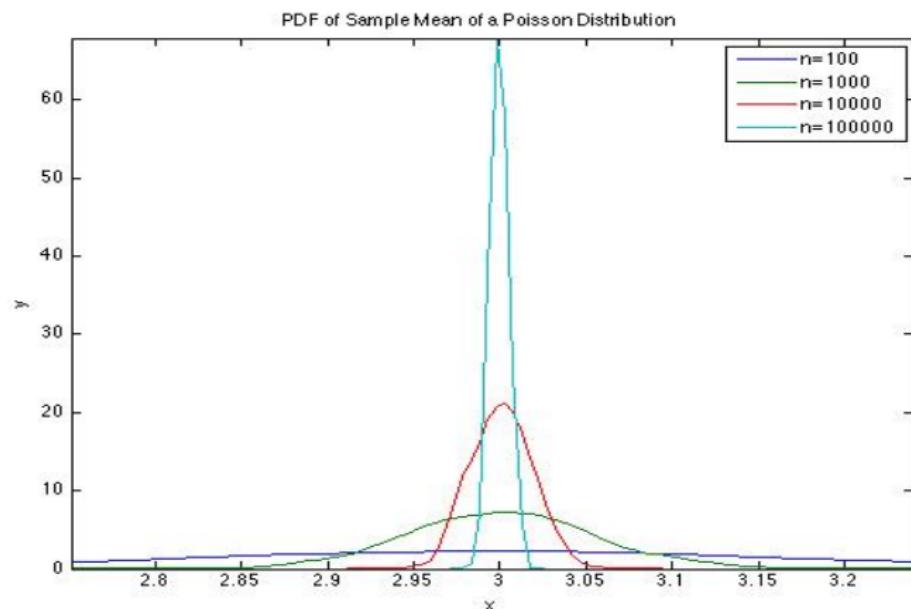
- Intuitive and non-rigorous proof of the LLN:
 - the expectation of the sample mean is the population mean;
 - the variance of the sample mean (i.e. average square distance to its expectation) decreases in the sample size.
 - therefore, as the sample size gets large, the sample mean will tend to be closer to the population mean.

Law of Large Numbers



Each PDF reflects the distribution of a 1000 sample means, each of them computed over a Poisson random sample of size n .

Law of Large Numbers



(This figure is the same as that in the previous slide. I have just modified the scale of the vertical axis for clarity)

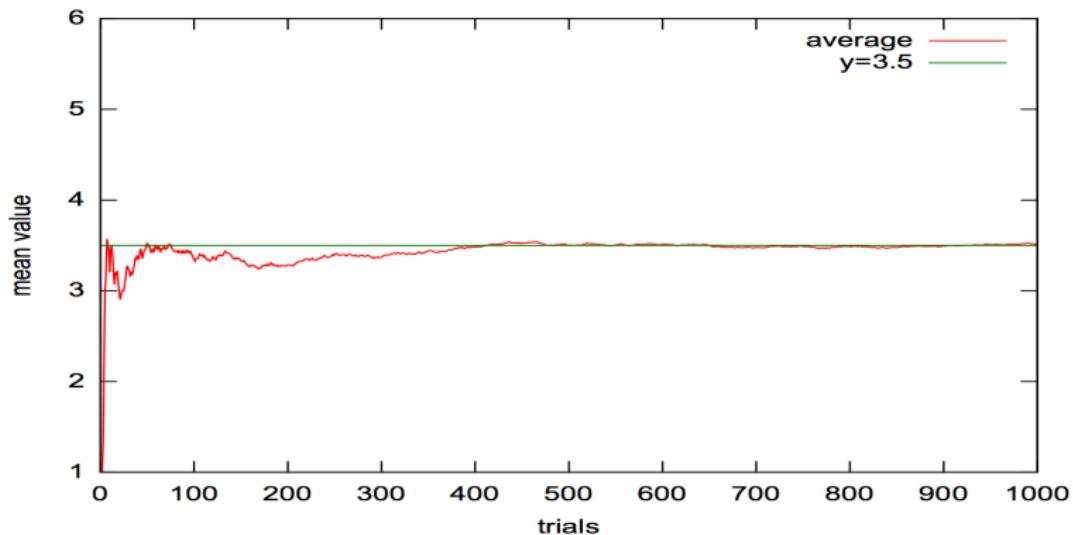
Law of Large Numbers

- How does the LLN differ from convergence of real numbers?
- What does convergence of real numbers mean?

A sequence of real numbers $\{a_n, n = 1, 2, \dots\}$ converges to some real number a if a_n gets arbitrarily close to a as n grows large.

- Eg: the sequence $\{a_n = a + 1/n, n = 1, 2, \dots\}$ converges to a as n increases.
- As n increases, we know *for sure* that a_n will be closer to a .
- In the case of the LLN, remember that \bar{Y}_n is a random variable no matter how large the sample is. Therefore, as the sample size n increases, it is possible that we observe a realization of \bar{Y}_n that is actually farther away from μ than the realization observed for a smaller n . The only thing we know is that *probability* of observing numbers that are far away from μ decreases *for sure* as n increases.

Law of Large Numbers



Sample average of the random variable capturing the outcome of tossing a die, as the number of tosses increases. Note that the realized sample average after 100 tosses is actually closer to 3.5 than the realized sample average after 200 tosses.

Central Limit Theorem

- Even though the sample average will get closer to the population mean as the sample size increases, for any finite sample (no matter how large this one is) the sample average will not be numerically equal to the population mean.
- Therefore, no matter how large the sample is, the sample mean will still have a distribution. You can see this in the figures in slides 27 and 28.
- What do we know about the shape of this distribution?
 - From property 2, we know that its mean is the population mean.
 - From property 3, we know that its variance is equal to the population variance divided by the sample size.
 - From property 4, we know that, **if the PDF in the population of interest is normal**, then the sample mean is also normally distributed.

Central Limit Theorem

- However, most of random variables are NOT normally distributed. What can we say about the shape of the distribution of the average of a random sample drawn from a distribution that is not normal?
- If the sample is “small”, we cannot say anything. Conversely, if the sample is “large”, then we know that the sampling distribution of the sample mean \bar{Y}_n is approximately normal, and the approximation gets better as the sample size n increases.
- This result is known as the Central Limit Theorem.
- **Independently of the distribution of Y in the population**, as long as n is sufficiently large, the sample mean, \bar{Y}_n , will be approximately normal with mean μ_Y and variance σ_Y^2/n :

$$\bar{Y}_n \approx \mathbb{N}(\mu_Y, \frac{\sigma_Y^2}{n}).$$

Central Limit Theorem

- Let's see the Central Limit Theorem in action.
- We will generate 5000 random samples of size n from three different random variables:
 - Poisson with parameter 1;
 - Benoulli with probability 0.5 of obtaining a 1;
 - Uniform(0,1);
- Once we have generated these 5000 random samples, we compute the sample average for each of them (i.e. we compute 5000 sample averages for each of the three distributions above) and plot the histogram that shows the distribution of these sample averages.
- For each of these distributions, we will write the exercise for different values of the sample size n : 1, 10, 100, 1000, 10000,
- As we increase the sample size n , this distribution of the sample average (computed across the 5000 random samples) will look closer to normal.
- You can find the STATA code that allows you to do this on blackboard.

Implications of LLN and CLT

- Household income is not distributed normally in America; it is skewed to the right.
 - The median household income in the United States is roughly \$51,900.
 - The average household income in the United States is roughly \$70,900.
- Suppose we take a random sample of 1,000 households (a “large” sample) and observe their annual income. What can we infer about this sample?
 - First, our best guess for the sample average is 70,900 (from slide 19).
 - Second, this guess is likely to be very accurate (from LLN). This is true only because our random sample is “large”.
 - Third, if we took multiple samples of 1,000 households, the distribution of the different sample means will look normal around the mean of \$70,900. This is true only because our random sample is “large”.
 - The household income distribution in the United States is very skewed, but the distribution of the sample means will not be skewed.

WWS 507c: Quantitative Analysis

Lecture 10: Inference about the Population Mean

Princeton University

November 4, 2014

PARAMETERS AND ESTIMATORS

Parameters and Estimators

- According to the U.S. Energy Administration, at least 10% of the world's remaining recoverable oil resources lie below the surface in deep water.
- To be economically viable, deep-water gas and oil reservoirs have to be big.
- Therefore, before engaging themselves into the construction of the infrastructure needed to exploit certain reservoirs, companies like ExxonMobil, Chevron, Royal Dutch Shell and BP first obtain **estimates** of the amount of gas and oil lying under the area of the ocean's surface that they are considering exploiting.
- These estimates are constructed by drilling in **randomly chosen** locations of the area of interest and inspecting whether there is oil or not under locations.

Parameters and Estimators

- This is what a random draw looks like...



Parameters and Estimators

- The same techniques used by oil companies to estimate the size of oil reservoirs may be used to estimate the proportion of earth covered by water.
- Imagine one were to divide the earth in 1 mile by 1 mile squares. The proportion of earth covered by water would be very well approximated by the fraction of these squares covered by water.
- Using the statistical terms we learned in Lecture 9, the **population of interest** is the set of all 1 mile by 1 mile squares in which we divide the earth.
- For each element of this population, we care about a random variable that takes value 1 if it is covered by water, and 0 if it is covered by land. Let's call this random variable as W .
- What we want to learn is the fraction of all the squares in the population of interest that are covered by water. This fraction is the so-called **parameter of interest**.

Parameters and Estimators

- In intuitive terms, a parameter is simply an unknown characteristic of the population of interest.
- In more formal terms, **a parameter is a function of the distribution of a random variable in the population of interest.**
- In our example, the parameter of interest is the expectation of the random variable W :

$$\theta = \mathbb{E}[W] = \frac{\text{number of squares covered by water in the population}}{\text{total number of squares in the population}}.$$

- In the case of deep-water drilling, the parameter of interest is the expectation of a random variable O that takes value 1 if there is oil underneath the ocean's surface and 0 otherwise.

$$\gamma = \mathbb{E}[O] = \frac{\text{number of squares with oil underneath in the population}}{\text{total number of squares in the population}}.$$

Parameters and Estimators

- In order to learn about the value that a parameter takes in the population of interest, we rarely have the option of measuring this parameter directly using data from all the individuals in the population.
- One would have to invest a very large amount of resources to:
 - drill every possible inch of the Gulf of Mexico in order to learn how much oil there is underneath;
 - ask every U.S. voter to learn about how much support Obama has among all the U.S. voters;
 - ask every U.S. resident between 18 and 65 years old about their employment status to learn about the U.S. unemployment rate;
 - measure the price of every economic transaction happening on U.S. soil in order to learn about the U.S. inflation rate;
- Instead, we generally use **random samples** from the population of interest in order to learn about parameters of this population.

Parameters and Estimators

- How can we extract a random sample that helps us obtain information about the proportion of earth covered by water?
- We would like to obtain a random sample of the 1 mile by 1 mile squares in which we have divided the earth.
- Using the following website

<http://planetaryjs.com/examples/rotating.html>

place your mouse over the moving earth, close your eyes, count until five and click the mouse as you open the eyes.

- In order to obtain a random sample of size 10, repeat this procedure 10 times.
- Each time $i = 1, \dots, 10$, create a random variable W_i that takes value 1 if the mouse is on water and 0 if it is on land.
- In this way, we can create a random sample of points on the globe:

$$\{W_1, W_2, \dots, W_{10}\}$$

Parameters and Estimators

- Once we have a random sample from the population of interest, how can we use it to learn something about the parameter of interest?
- We compute the so-called **estimators**.
- In our case, the parameter of interest is the expectation of W or fraction of 1 mile by 1 mile squares in which we divide the earth that are covered by water.
- Possible estimators that we can construct using our random sample are:
 - sample average: $\bar{W} = (1/10) \sum_{i=1}^{10} W_i$
 - the median: $med(W)$
 - the square root of the 1st observation multiplied by the 4th one: $\sqrt{W_1} \times W_4$
- If, for example, our realized random sample was:

$$\{W_1, W_2, \dots, W_{10}\} = \{0, 0, 0, 0, 0, 0, 0, 1, 1, 1\}$$

then our estimators would be:

- sample average: 0.7
- median: 0.5
- the square root of the 1st observation multiplied by the 4th one: 0

Parameters and Estimators

- In informal terms, an estimator of a parameter is a guess that we make about this parameter using the information contained in a sample.
- In formal terms, an **estimator of a parameter** θ , denoted as $\hat{\theta}$, is a **function of a random sample** of the population of interest, $\{W_1, W_2, \dots, W_{10}\}$, that is used to provide information about θ .
- Note that all estimators are random variables: different random samples from the same population of interest will generate different values of $\hat{\theta}$.
- Therefore, estimators will have CDF, PDF, expectation, variance, etc.
- The particular value or realization that an estimator takes in one particular random sample is called **estimate**.
- Example: the sample average is one possible estimator. If the observed random sample is

$$\{0, 0, 0, 0, 0, 0, 0, 1, 1, 1\},$$

then the estimate of the sample average is 0.7.

Parameters and Estimators

- As we saw in slide 9, one can define different estimators of a single parameter.
- Different estimators take different estimates, even if we apply them to the same random sample.
- Some estimators are better than others.
- The ideal estimator of a parameter θ would be an estimator $\hat{\theta}$ such that
 - independently of the value of θ in the population of interest; and
 - in every possible random sample from the population of interest,takes a value identical to the value of θ in the population of interest.
- If the true share of earth covered by water is 0.7, we would like our estimator to take the value 0.7 independently of whether our random sample contains 10 squares covered by water or 5 squares covered by water.
- This ideal estimator does not exist.

Parameters and Estimators

- This is the dilemma we face when choosing a good estimator:
- On the one side, we want our estimators to vary with every possible realization of a random sample from the population of interest. The reason is that this sample is all the information we have about the population.
 - If we take a random sample of a 1,000 individuals from the U.S. active population and they are all unemployed, we want our estimator of the unemployment rate in the U.S. to be high.
- On the other side, a particular realization of a random sample might be very uninformative about the population of interest and, in this case, our estimates might be very far away from the true value of the parameter of interest.
 - If we take a random sample of a 1,000 individuals from the U.S. active population and the unemployment rate is 1%, it is still possible that all the 1,000 individuals in our sample are unemployed.

Parameters and Unbiased Estimators

- Given that the ideal estimator does not exist, which properties would we like our estimators to have?
- First**, we would like that the average of the estimates generated by our estimator across all the possible **random samples** that we can take from the population of interest is equal to the parameter we are trying to estimate. If an estimator has this property, we say that the estimator is **unbiased**.
- If an estimator is not unbiased, then it must be **biased**.
- How can we write this in mathematical terms? Using the definition of expectation, we can write that an estimator $\hat{\theta}$ is an unbiased estimator of a parameter θ if $\mathbb{E}[\hat{\theta}] = \theta$, where the expectation is over all possible random samples from the population of interest.
- On the contrary, $\hat{\theta}$ will be a biased estimator of a parameter θ if $\mathbb{E}[\hat{\theta}] \neq \theta$.
- The difference between the expectation of an estimator and the parameter it aims to estimate is the **bias**: $\text{bias} = \mathbb{E}[\hat{\theta} - \theta]$.

Parameters and Unbiased Estimators

- In intuitive terms, what does unbiasedness mean?
- Imagine that we are trying to estimate the unemployment rate in the U.S. using samples of size 2. Therefore, the parameter of interest, θ , is the fraction of all potential workers in the U.S. that are unemployed (i.e. 0.059)
- There are many samples of size 2 that we can extract from the U.S. active population: {John and Mary}, {you and Mary}, {you and you}, etc.
- For each of these samples of size 2, we can compute the value of our preferred estimator. For example, let's define our estimator, $\hat{\theta}$, as the sample average of a random variable that takes value 1 if an individual is unemployed.
- If both John and Mary are unemployed, then $\hat{\theta}(\{\text{John and Mary}\}) = 0$.
- If Peter is unemployed but Sarah is employed, $\hat{\theta}(\{\text{Peter and Sarah}\}) = 0.5$.
- $\hat{\theta}$ is unbiased if the average of $\hat{\theta}$ across the millions and millions of possible random samples of size 2 from the U.S. active population is equal to 0.059 (i.e. the actual unemployment rate in the population of interest).

Parameters and Consistent Estimators

- The **second** property that we would like our estimator to have is consistency.
- An estimator is **consistent** if the probability that it takes a value that differs from the value taken by the parameter of interest in any given number k decreases as the sample size increases.
- If an estimator is not consistent, then it is **inconsistent**.
- How can we write this in mathematical terms?

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > k) = 0,$$

where the probability is over all the possible random samples from the population of interest.

- If $\hat{\theta}$ is a consistent estimator for θ , we can also write

$$\hat{\theta} \xrightarrow{P} \theta$$

Parameters and Consistent Estimators

- In intuitive terms, what does consistency mean?
- In very informal terms, it just means that, as the random sample increases in size, the estimator is more likely to be close to the parameter of interest.
- In a little bit more precise terms...
- Imagine that we are trying to estimate the unemployment rate in the U.S. ($\theta = 0.059$) using either samples of size 100 or samples of size 1000.
- There are many random samples of size 100 that we can extract from the U.S. population. For each of them we compute the sample average of a random variable that takes value 1 if an individual is unemployed. We denote the value that we obtain for a sample i of size 100 as $\hat{\theta}_{100i}$.
- If we take many many many random samples of size 100 and compute the fraction of them for which $\hat{\theta}$ differs from 0.059 in more than, for e.g. 3%, we would be computing

$$P(|\hat{\theta}_{100} - 0.059| > 0.03) = P(0.029 < \hat{\theta}_{100} < 0.089)$$

Parameters and Consistent Estimators

- If we take many many many random samples of size 1000 and compute the fraction of them for which $\hat{\theta}$ differs from 0.059 in more than, for e.g. 3%, we would be computing

$$P(|\hat{\theta}_{1000} - 0.059| > 0.03) = P(0.029 < \hat{\theta}_{1000} < 0.089)$$

- If the estimator $\hat{\theta}$ is consistent, then, when computed on larger samples, it is more likely that its distribution is concentrated around the true value of the parameter:

$$P(0.029 < \hat{\theta}_{1000} < 0.089) > P(0.029 < \hat{\theta}_{100} < 0.089)$$

- Even more, if $\hat{\theta}$ is consistent, this would be true for any two sample sizes that we want to compare and for any distance to the true parameter vector that we want to choose instead of the 3% picked in this example.

Parameters and Efficient Estimators

- The third property that we would like our estimator to have is **low variance**.
- We say that an estimator $\hat{\theta}_{1n}$ is **more efficient than** an alternative estimator $\hat{\theta}_{2n}$ if

$$\mathbb{V}[\hat{\theta}_{1n}] \leq \mathbb{V}[\hat{\theta}_{2n}],$$

where the variance is across all possible random samples from the population of interest of size n .

Estimators and Mean Squared Errors

- Imagine that an estimator $\hat{\theta}_1$ is more efficient than an alternative estimator $\hat{\theta}_2$ but the bias of $\hat{\theta}_1$ is larger in absolute value than that of $\hat{\theta}_2$:

$$\mathbb{V}[\hat{\theta}_1] < \mathbb{V}[\hat{\theta}_2] \quad \text{but} \quad |\text{bias}(\hat{\theta}_1)| > |\text{bias}(\hat{\theta}_2)|$$

Which of the two estimators is preferred? The usual criterium is to choose the estimator that minimizes the **mean squared error**:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

- Note that we can rewrite the MSE as a function of more familiar elements:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta})) + (\mathbb{E}(\hat{\theta}) - \theta)]^2 \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2 + (\mathbb{E}(\hat{\theta}) - \theta)^2 + 2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + \mathbb{E}[(\mathbb{E}(\hat{\theta}) - \theta)^2] + \mathbb{E}[2(\hat{\theta} - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta)] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2] + (\mathbb{E}(\hat{\theta}) - \theta)^2 + 2(\mathbb{E}(\hat{\theta}) - \mathbb{E}(\hat{\theta}))(\mathbb{E}(\hat{\theta}) - \theta) \\ &= \mathbb{V}(\hat{\theta}) + [\text{bias}(\hat{\theta})]^2\end{aligned}$$

ESTIMATION OF THE POPULATION MEAN: SAMPLE AVERAGE

Estimation of the Population Mean

- As an estimator of the population mean, the sample average of a random sample from the population of interest has the following properties:
 - it is unbiased;
 - it is consistent;
 - it has lower variance than any other unbiased **linear** estimator;
- An estimator is said to be **linear** as long as, for any random sample of size n of a random variable Y

$$\{Y_1, Y_2, \dots, Y_n\}$$

we can write our estimator as

$$\hat{\theta} = \omega_1 Y_1 + \omega_2 Y_2 + \dots + \omega_n Y_n.$$

where

$$\{\omega_1, \omega_2, \dots, \omega_n\}$$

are just a set of real numbers.

Estimation of the Population Mean

- As long as we have a random sample of a variable Y , the sample average, \bar{Y} , is an unbiased estimator for the population mean, μ_Y .
- Proof:

$$\mathbb{E}[\bar{Y}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n Y_i\right] = \frac{1}{n} n \mu_Y = \mu_Y$$

- Is the sample average the only unbiased estimator of the population mean? No. Imagine that we define a general estimator $\hat{\theta}$ as

$$\hat{\theta}(\{\omega_i\}) = \sum_{i=1}^n \omega_i Y_i,$$

where $\{\omega_i, i = 1, \dots, n\}$ is a set of real numbers. Any estimator of this kind is unbiased as long as $\sum_{i=1}^n \omega_i = 1$.

Estimation of the Population Mean

- Among all the different estimators $\hat{\theta}(\{\omega_i\})$ such that $\sum_{i=1}^n \omega_i = 1$, which one has the lowest variance?

$$\begin{aligned}\mathbb{V}(\hat{\theta}(\{\omega_i\})) &= \mathbb{E}[(\hat{\theta}(\{\omega_i\}) - \mathbb{E}[\hat{\theta}(\{\omega_i\})])^2] = \\ \mathbb{E}\left[\left(\sum_{i=1}^n \omega_i Y_i - \mathbb{E}\left[\sum_{i=1}^n \omega_i Y_i\right]\right)^2\right] &= \mathbb{E}\left[\left(\sum_{i=1}^n \omega_i (Y_i - \mathbb{E}[Y_i])\right)^2\right] = \\ \mathbb{E}\left[\sum_{i=1}^n \omega_i^2 (Y_i - \mathbb{E}[Y_i])^2 + 2 \sum_{i=1}^n \sum_{i' \neq i} \omega_i (Y_i - \mathbb{E}[Y_i]) \omega_{i'} (Y_{i'} - \mathbb{E}[Y_{i'}])\right] &= \\ \sum_{i=1}^n \omega_i^2 \mathbb{E}[(Y_i - \mathbb{E}[Y_i])^2] + 2 \sum_{i=1}^n \sum_{i' \neq i} \omega_i \omega_{i'} \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_{i'} - \mathbb{E}[Y_{i'}])] &= \\ \sigma_Y^2 \sum_{i=1}^n \omega_i^2. \end{aligned}$$

Estimation of the Population Mean

- Therefore, among all the linear estimators $\hat{\theta}(\{\omega_i\})$ of the population mean that are unbiased (i.e. such that $\sum_{i=1}^n \omega_i = 1$), the one that has the smallest variance is the one such that $\sum_{i=1}^n \omega_i^2$ is minimized.
- If you solve the problem

$$\min_{\{\omega_i\}} \sum_{i=1}^n \omega_i^2 \quad \text{s.t.} \quad \sum_{i=1}^n \omega_i = 1,$$

you will find out that the solution is $\omega_i = \frac{1}{n} \forall i$. And note that

$$\hat{\theta}(\{\omega_i = \frac{1}{n}\}) = \sum_{i=1}^n \omega_i Y_i = \sum_{i=1}^n \frac{1}{n} Y_i = \bar{Y}.$$

- Therefore, as long as $\{Y_i\}$ are *i.i.d*, the sample average, $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$, is the **best linear unbiased estimator of the population mean** μ_Y .

Estimation of the Population Mean

- The sample mean \bar{Y} is a consistent estimator of the population mean μ_Y .
- An estimator that simply picks the first observation of a random sample,

$$\hat{\mu}_Y = Y_1$$

is unbiased but not consistent.

- An estimator that adds $1/n$ to the sample average, n being the size of the random sample,

$$\hat{\mu}_Y = \bar{Y} + \frac{1}{n}$$

is biased but consistent.

ESTIMATION OF THE POPULATION VARIANCE: SAMPLE VARIANCE

Estimation of the Population Variance

- Imagine you have access to a **random sample** of U.S. workers and you want to use data on their wages to learn about U.S. wage inequality.
- One measure of income inequality that might be useful to you is the variance:

$$\sigma_Y^2 = \mathbb{E}[(Y - \mu_Y)^2].$$

- Which other measures of income inequality you can think of?
- How can we use a random sample to learn about the population variance?
- A natural estimator of the population variance is the **sample variance**:

$$\hat{\sigma}_Y^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

- Given access to a **random sample** from the population of interest, which are the properties of $\hat{\sigma}_Y^2$ as estimator of σ_Y^2 ?
 - ① Is it unbiased?
 - ② Is it more efficient than other estimators?
 - ③ Is it consistent?

Estimation of the Population Variance

- Is the sample variance an unbiased estimator of the population variance?

$$\mathbb{E}[\hat{\sigma}_Y^2] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})^2\right] = \frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n ((Y_i - \mu_Y) - (\bar{Y} - \mu_Y))^2\right] =$$

$$\frac{1}{n} \mathbb{E}\left[\sum_{i=1}^n ((Y_i - \mu_Y)^2 + (\bar{Y} - \mu_Y)^2 - 2(Y_i - \mu_Y)(\bar{Y} - \mu_Y))\right] =$$

$$\frac{1}{n} \sum_{i=1}^n \{\mathbb{E}[(Y_i - \mu_Y)^2]\} + \frac{1}{n} \sum_{i=1}^n \{\mathbb{E}[(\bar{Y} - \mu_Y)^2]\} - \frac{2}{n} \sum_{i=1}^n \mathbb{E}[(Y_i - \mu_Y)(\bar{Y} - \mu_Y)] =$$

$$\frac{1}{n} n \sigma_Y^2 + \frac{1}{n} n \frac{\sigma_Y^2}{n} - \frac{2}{n} n \frac{\sigma_Y^2}{n} = \sigma_Y^2 + \frac{\sigma_Y^2}{n} - 2 \frac{\sigma_Y^2}{n} = \left(\frac{n-1}{n}\right) \sigma_Y^2$$

where we have used that

$$\mathbb{E}[(Y_i - \mu_Y)(\bar{Y} - \mu_Y)] = \mathbb{E}\left[(Y_i - \mu_Y)\left(\frac{1}{n} \sum_{j=1}^n Y_j - \frac{n}{n} \mu_Y\right)\right] =$$

$$\frac{1}{n} \mathbb{E}\left[(Y_i - \mu_Y)\left(\sum_{j=1}^n (Y_j - \mu_Y)\right)\right] = \frac{1}{n} \sum_{j=1}^n \mathbb{E}[(Y_i - \mu_Y)(Y_j - \mu_Y)] = \frac{1}{n} \mathbb{E}[(Y_i - \mu_Y)^2]$$

Estimation of the Population Variance

- Therefore, even if our sample is random, the sample variance is NOT an unbiased estimator of the population variance.
- The bias is equal to

$$\mathbb{E}[\hat{\sigma}_Y^2] - \sigma_Y^2 = -\frac{1}{n}\sigma_Y^2$$

- The bias disappears as the sample size gets large.
- We have not computed it in these slides, but the variance of the sample variance also gets smaller and smaller as the sample size increases.
- The sample variance is a consistent estimator of the population variance.
- It is easy to construct an estimator that is unbiased:

$$s_Y^2 = \left(\frac{n}{n-1} \right) \hat{\sigma}_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

- In practice, it does not matter whether we use $\hat{\sigma}_Y^2$ or s_Y^2 to estimate σ_Y^2 .

ESTIMATION OF STANDARD DEVIATION, COVARIANCE AND CORRELATION COEFFICIENT

Other Estimators

- As an estimator of the **standard deviation**, σ_Y , we will use the square-root of the sample variance:

$$\hat{\sigma}_Y = \sqrt{\hat{\sigma}_Y^2}$$

- As an estimator of the **covariance** between two variables,

$$\sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)],$$

we will use the sample covariance,

$$\hat{\sigma}_{XY} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}).$$

- As an estimator for the population **correlation coefficient**, ρ_{XY} , we will use the sample correlation coefficient:

$$r_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X \hat{\sigma}_Y}.$$

- If the sample is random, all these estimators are consistent.**

HYPOTHESIS TESTS

Estimation vs. Hypothesis Tests

- Conditional on having access to a random sample from the population of interest, knowing the properties of estimators is helpful for finding the best “guess” about a characteristic of the population.
- Knowing the properties of estimators and data from a random sample will help you formulate an educated “guess” as the answer to questions like:
 - What percentage of Americans do you think are Christian?
 - What percentage of Americans do you think are unemployed?
 - What percentage of Americans do you think are immigrants?
 - Out of every 100 eligible voters, how many do you think voted in the last presidential election?
- If you want to know the answer, go to

https://www.ipsos-mori.com/_assets/perceptionquiz/index.html

Estimation vs. Hypothesis Tests

- However, in many occasions, we do not want to use the information contained in an estimator to make guesses about a characteristic of the distribution of a variable in the population of interest, but rather to answer a true/false question about this characteristic.
- Examples of true/false or yes/no questions we might care about:
 - Is the percentage of Americans who are immigrants 6%?
 - Will Nevada's Ballot Question 3 be approved? (If approved, it would impose a 2% tax on business profits, with the revenue dedicated to schools).
 - Are mean rents for a one bedroom apartment the same in NYC and in SF?
 - Are voting-age individuals reached by get-out-the-vote phone campaigns more likely to vote than non-reached voters?
 - Are smokers as likely as non-smokers to suffer cancer?
- In all these questions, we are asking whether a **quantitative** statement about the population of interest is true or false.

Examples of Tests

- In the language of statistics, the statement about the population of interest whose truth or falseness we are trying to determine is called **null hypothesis**.
- The statement that becomes true if the null hypothesis is false is the **alternative hypothesis**.
- In the example “*Is the percentage of Americans who are immigrants 6%?*”, the null hypothesis is

$$H_0 : \mu_Y = 6\%$$

and the alternative hypothesis is

$$H_1 : \mu_Y \neq 6\%,$$

where μ_Y is the mean in the population of Americans of a Bernoulli random variable taking value 1 for an individual if she is an immigrant.

- This is an example of a **hypothesis test about the mean of a population**. The alternative hypothesis is said to be a **two-sided alternative**.

Examples of Tests

- In the example “*Will Nevada’s Ballot Question 3 be approved?*”, the null hypothesis is

$$H_0 : \mu_Y \geq 50\%,$$

and the alternative hypothesis is

$$H_1 : \mu_Y \leq 50\%,$$

where μ_Y is the mean across the population of actual voters of a Bernoulli random variable taking value 1 for an individual if she votes in favor of Ballot Question 3.

- This is also an example of a **hypothesis test about the mean of a population**. However, in this case, the alternative hypothesis is said to be **one-sided**.

Examples of Tests

- In the example “Are mean rents for a one bedroom apartment the same in NYC and in SF?”, the null hypothesis is

$$H_0 : \mu_{NYC} = \mu_{SF},$$

and the alternative hypothesis is

$$H_1 : \mu_{NYC} \neq \mu_{SF},$$

where μ_A is the mean of a random variable capturing the rental price across the population of 1-bedroom apartments available for rent in A .

- This is an example of a **test of equality of means across two populations**. The alternative hypothesis is said to be a **two-sided alternative**.

Examples of Tests

- The example

- Are smokers as likely as non-smokers to suffer cancer?

also implies tests of equality of means across two populations and the alternative hypotheses are two-sided:

$$H_0 : \mu_S = \mu_{NS} \quad \text{and} \quad H_1 : \mu_S \neq \mu_{NS},$$

where μ_S denotes the mean in the population of smokers of a Bernoulli random variable taking value 1 if a person has suffered from cancer, and μ_{NS} denotes the mean of an identical variable in the population of non-smokers.

- Which is the null and the alternative hypothesis in the example “*Are voting-age individuals reached by get-out-the-vote phone campaigns more likely to vote than non-reached voters?*”

Difference in Means vs. Causal Relationships

- It is **very important** to realize that testing the equality of means across two populations does not imply testing a causal relationship.
- In the hypothesis test *“Are voting-age individuals reached by get-out-the-vote phone campaigns more likely to vote than non-reached voters?”*, it is perfectly possible that the null hypothesis

$$H_0 : \mu_R > \mu_{NR},$$

is true in the population and that the get-out-the-vote phone campaigns have no impact on determining whether voting-age individuals actually vote. For example, this would be true if the campaign happens to reach mostly individuals that would have gone to vote even if they had not been reached.

- μ_R denotes the mean across the population of potential voters reached by get-out-the-vote phone campaign if a Bernoulli random variable equal 1 of a person actually goes to vote. Analogously for μ_{NR} .

Difference in Means vs. Causal Relationships

- The opposite is also possible.
- It might be true that

$$H_0 : \mu_R < \mu_{NR},$$

and that the get-out-the-vote phone campaigns have a positive impact on the probability that a voting-age individual goes to vote. This could happen if the get-out-the-vote phone campaigns target mostly individuals that, for other reasons, are less likely to vote.

- Summary: imagine we split the population of interest into two subpopulations according to the value of some random variable X (e.g. smoking vs. not smoking), and we observe that the mean of some random variable Y (e.g. dummy for suffering cancer) is higher in one of these subpopulations, it would be **incorrect** to conclude from this evidence **alone** that X causes Y .

TESTING HYPOTHESIS ABOUT THE MEAN OF A POPULATION

Intuition Behind Hypothesis Testing: Two-Sided Test

- Given a particular null and alternative hypotheses, how can we use the information contained in a random sample from the population of interest to conclude which of the two hypotheses is true?
- To fix ideas, assume for now that we are trying to perform a two-sided test

$$H_0 : \mu_Y = 6\% \quad \text{vs.} \quad H_1 : \mu_Y \neq 6\%,$$

where μ_Y denotes the probability of being an immigrant.

- Assume also that the only information you have is a random sample from the population of interest. As we saw before in this Lecture 10, the sample average of Y , \bar{Y} , is the best summary of the information contained in this random sample about the population mean μ_Y .
- However, it is possible that $\mu_Y = 6\%$ and, in our particular random sample, $\bar{Y} = 99\%$. Or that we observe $\bar{Y} = 6\%$ even though $\mu_Y = 99\%$.

Intuition Behind Hypothesis Testing: Two-Sided Test

- Therefore, using the information contained in a random sample only, we cannot prove right or wrong any null hypothesis. We **cannot** say anything about the population **with full certainty**.
- At the same time, we know that the sample average of a random sample is normally distributed around the population mean. Given that the normal distribution is bell-shaped, this implies that the probability that we observe a sample average, \bar{Y} , that differs from the population mean, μ_Y , in more than any given distance d (e.g. 3 percentage points) decreases as d increases:

$$Pr(|\bar{Y} - \mu_Y| > d) \text{ decreases as } d \text{ increases}$$

- So, the information contained in a random sample does not allow us to reject a null hypothesis with full certainty, but this does not mean that this random sample does not contain any information at all about the null hypothesis.

TESTING HYPOTHESIS ABOUT THE MEAN OF A POPULATION: COMPUTING THE P-VALUE

Intuition Behind Hypothesis Testing: Two-Sided Test

- If we are interested in testing

$$H_0 : \mu_Y = \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_0,$$

where μ_0 is a particular number (e.g. 6%), how can we exploit the info contained in a random sample of Y from the population of interest?

- Imagine that, in our particular random sample, we observe a particular value of the sample average \bar{Y}^{obs} . We will then report the probability that, **if the null hypothesis were to be true**, we observe a sample average that differs from the population mean implied by the null hypothesis, μ_0 , in a distance that is as large as that observed between μ_0 and \bar{Y}^{obs} . This probability is the **p-value for a two-sided test**:

$$\text{p-value} = P_0(|\bar{Y} - \mu_0| > |\bar{Y}^{obs} - \mu_0|)$$

- If the p-value is small, then the information contained in the observed sample is telling us that it is unlikely that the null hypothesis is true.

How do we Compute the p-value in a Two-Sided Test?

- As the definition of p-value shows, a requisite to compute the p-value of

$$H_0 : \mu_Y = \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_0,$$

is to know the distribution of the sample average \bar{Y} conditional on $\mu_Y = \mu_0$.

- Using the CLT, we know that, in a random sample of size n , with n large,

$$\bar{Y} \approx \mathbb{N}(\mu_Y, \frac{\sigma_Y^2}{n}).$$

Therefore, conditional on the null hypothesis $H_0 : \mu_Y = \mu_0$ being true, the distribution of the sample average is

$$\bar{Y} \approx \mathbb{N}(\mu_0, \frac{\sigma_Y^2}{n}).$$

How do we Compute the p-value in a Two-Sided Test?

- Therefore, the p-value is

$$\begin{aligned} P_0(|\bar{Y} - \mu_0| > |\bar{Y}^{obs} - \mu_0|) &= \\ P_0(\bar{Y} - \mu_0 > |\bar{Y}^{obs} - \mu_0|) + P_0(\mu_0 - \bar{Y} > |\bar{Y}^{obs} - \mu_0|) &= \\ P_0(\bar{Y} - \mu_0 > |\bar{Y}^{obs} - \mu_0|) + P_0(\bar{Y} - \mu_0 < -|\bar{Y}^{obs} - \mu_0|) &= \\ P_0\left(\frac{\bar{Y} - \mu_0}{\sigma_Y/\sqrt{n}} > \frac{|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}\right) + P_0\left(\frac{\bar{Y} - \mu_0}{\sigma_Y/\sqrt{n}} < \frac{-|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}\right) &= \\ P(Z > \frac{|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}) + P(Z < \frac{-|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}) &= \\ 2 \times P(Z > \frac{|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}) &= \\ 2 \times \left(1 - \Phi\left(\frac{|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}\right)\right). \end{aligned}$$

- This expression for the p-value assumes that σ_Y is known. If σ_Y unknown, we just need to substitute σ_Y by $\hat{\sigma}_Y$ and everything else stays the same.

How do we Compute the p-value in a Two-Sided Test?

- Example.
- Imagine that you are trying to test

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 0,$$

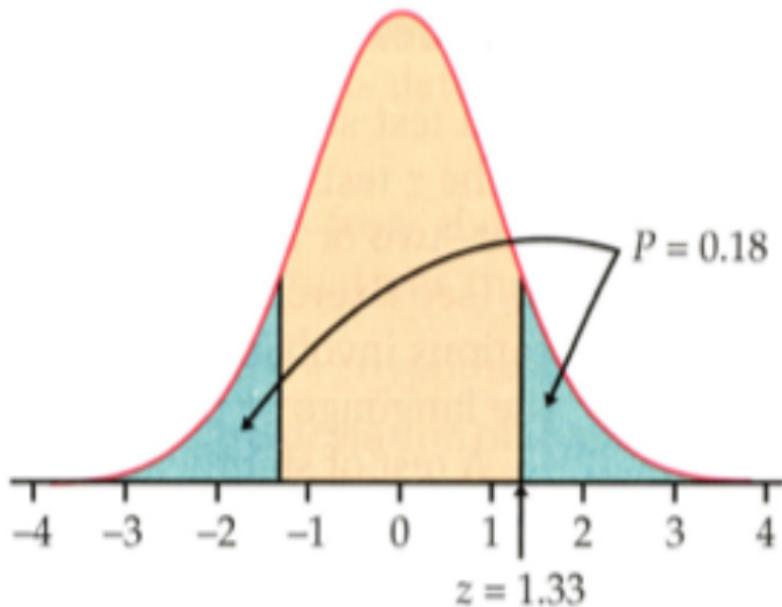
and you have a random sample such that $\bar{Y} = 126$ and $\text{var}(\bar{Y}) = 95^2$.

- Using the previous formula, the p-value is:

$$\begin{aligned} 2 \times \left(1 - \Phi \left(\frac{|\bar{Y}^{obs} - \mu_0|}{\sigma_{\bar{Y}} / \sqrt{n}} \right) \right) &= 2 \times \left(1 - \Phi \left(\frac{|126 - 0|}{95} \right) \right) = \\ 2 \times (1 - \Phi(1.33)) &= 2(1 - 0.9082) = 2 \times 0.0918 = 0.1836. \end{aligned}$$

How do we Compute the p-value in a Two-Sided Test?

- The p-value is equal to the area under the density function painted in green.



How do we Compute the p-value in a Two-Sided Test?

- Example.
- We want to test

$$H_0 : \mu_Y = 1,500 \quad \text{vs.} \quad H_1 : \mu_Y \neq 1,500,$$

where Y denotes the rental price of a 1-bedroom apartment. Assume that $\sigma_Y = 500$ and that, in a random sample of size 100, we observe a sample average equal to 1,600. The p-value will be equal to:

$$2\left(1 - \Phi\left(\frac{|\bar{Y}^{obs} - \mu_0|}{\sigma_Y/\sqrt{n}}\right)\right) = 2\left(1 - \Phi\left(\frac{|1600 - 1500|}{500/\sqrt{100}}\right)\right) = 2(1 - \Phi(2))$$

Imagine that we actually do not know the variance of Y in the population of interest, but, in our sample, we estimate the standard deviation of Y to be 550. In this case, we can compute the p-value as:

$$2\left(1 - \Phi\left(\frac{|\bar{Y}^{obs} - \mu_0|}{\hat{\sigma}_Y/\sqrt{n}}\right)\right) = 2\left(1 - \Phi\left(\frac{|1600 - 1500|}{550/\sqrt{100}}\right)\right).$$

Intuition Behind Hypothesis Testing: One-Sided Test

- If we are interested in testing

$$H_0 : \mu_Y \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y < \mu_0,$$

where μ_0 is a particular number, how can we exploit the info contained in a random sample of Y from the population of interest?

- Imagine that, in our particular random sample, we observe a particular value of the sample average \bar{Y}^{obs} . We will then report the probability that, **if the null hypothesis were to be true**, we observe a sample average that is smaller than the observed sample average \bar{Y}^{obs} . This probability is the **p-value for a one-sided test**:

$$\text{p-value} = P_0(\bar{Y} < \bar{Y}^{obs})$$

- If the p-value is small, then the information contained in the observed sample is telling us that it is unlikely that the null hypothesis is true.

How do we Compute the p-value in a One-Sided Test?

- Therefore, the p-value for the one-sided test in the previous slide is:

$$P_0(\bar{Y} < \bar{Y}^{obs}) = P_0\left(\frac{\bar{Y} - \mu_0}{\sigma_Y/\sqrt{n}} < \frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y/\sqrt{n}}\right) = P(Z < \frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y/\sqrt{n}}) = \Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y/\sqrt{n}}\right).$$

- How would you compute the p-value associated to the following test:

$$H_0: \mu_Y \leq \mu_0 \quad \text{vs.} \quad H_1: \mu_Y > \mu_0,$$

where μ_0 is a particular number?

$$P_0(\bar{Y} > \bar{Y}^{obs}) = P_0\left(\frac{\bar{Y} - \mu_0}{\sigma_Y/\sqrt{n}} > \frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y/\sqrt{n}}\right) = P(Z > \frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y/\sqrt{n}}) = 1 - \Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y/\sqrt{n}}\right).$$

How do we Compute the p-value in a One-Sided Test?

- Imagine that you have access to data on a random sample of voters in Nevada and, in your sample of size 100, 49% of the voters voted in favor of approving Question 3. Would you reject the null hypothesis that

$$H_0 : \mu_Y \geq 50\%,$$

in favor of the alternative hypothesis $H_1 : \mu_Y < 50\% ?$

- In this case

$$\bar{Y} = 0.49 \quad \text{and} \quad \text{var}(\bar{Y}) = \frac{0.5(1 - 0.5)}{100} = 0.0025$$

so

$$\Phi\left(\frac{0.49 - 0.50}{\sqrt{0.0025}}\right) = \Phi(-0.2) = 0.4207.$$

- Even if the true underlying probability is 0.5, in 42% of the random samples of size a 100 you will obtain a sample average smaller than 0.49.

TESTING HYPOTHESIS ABOUT THE MEAN OF A POPULATION: REJECTING (OR NOT) A NULL HYPOTHESIS.

Rejecting (or Not) a Null Hypothesis

- Remember that we motivated hypothesis testing as a procedure that helps us make up our minds about whether a statement about a population of interest contained in a null hypothesis is true or false.
- So far we have only seen how to use the information in a random sample to compute the p-value associated to a null hypothesis that contains a statement about the mean of a random variable. How do we go from the p-value to a yes/no answer?
- In informal terms, we will reject the null hypothesis if the distance between the observed sample average and the population mean implied by the null hypothesis is “too large”.
- What does “too large” mean? It means that it is “too unlikely”, **if the null hypothesis were to be true**, that we observe a sample average that differs from the population mean implied by the null hypothesis, μ_0 , in a distance that is as large as that observed between μ_0 and \bar{Y}^{obs} .

Rejecting (or Not) a Null Hypothesis

- In other words, given an observed random sample and its average, we will reject a null hypothesis if the p-value is “too small”.
- When is the p-value “too small”? In statistics, it is common to use 1%, 5% or 10% thresholds. This threshold is called the **significance level** of a test and it is usually denoted as α .
- Let’s revisit our example in slide 53. You have access to data on a random sample of voters in Nevada and, in your sample of size 100, 49% of the voters voted in favor of approving Question 3. Given a significance level $\alpha = 0.1$, would you reject the null hypothesis that

$$H_0 : \mu_Y \geq 50\%,$$

in favor of the alternative hypothesis $H_1 : \mu_Y < 50\% ?$

- From slide 53, the p-value is 0.4207. As **p-value > significance level**, we do **not** reject.

Rejecting (or Not) a Null Hypothesis

- Would you have rejected if the sample average had been 40%?
- In this case, the p-value would be

$$\Phi\left(\frac{0.40 - 0.50}{\sqrt{0.0025}}\right) = \Phi(-2) = 0.0228,$$

and, therefore, if your desired significance level is 10%, you would reject the null hypothesis that the share of voters in the population supporting the approval of Question 3 is larger than 50%. If your desired significance level was 1%, then you would not reject the null hypothesis even if the share of “yes” votes observed in the random sample of size 100 was 40%.

- The smaller the significance level is, the less likely it is that you reject the null hypothesis.

Rejecting (or Not) a Null Hypothesis

- If the only information we have is that contained in a random sample, then we can never be certain about whether H_0 is true or not. Therefore, whenever we reject or we fail to reject a null hypothesis, we can always make errors.
- There are two types of errors we might make
 - type I error: we reject the null hypothesis when it is true.
 - type II error: we fail to reject the null hypothesis when it is not true.
- The significance level is precisely the probability of making a type I error.
- In order to understand this, remember that the p-value is the probability, **conditional on the null hypothesis being true**, of observing a value of the sample average, \bar{Y} , as far from the value of the mean implied by the null hypothesis, μ_0 , as the observed sample average, \bar{Y}^{obs} . Given that we reject a null hypothesis whenever the p-value is smaller than the significance level, then this significance level is precisely the probability of rejecting a null hypothesis **conditional on this null hypothesis being true**.
- The probability that we reject H_0 when it is false is called **power**.

Rejecting (or Not) a Null Hypothesis

- Example.
- Imagine that the mean of a distribution of a random variable Y can only take two values: 0 and 1. Therefore, you want to test:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu = 1.$$

Imagine also that your desired significance level is 0.05. You also know that the variance of Y in the population is equal to 4. If you observe a random sample of size 25, how high the observed sample average \bar{Y} must be for you to reject H_0 in favor of H_1 .

- Note that this is a one-sided test. If \bar{Y} is way below 0, we do not reject the null. We only reject the null if \bar{Y} is too large. Given our chosen significance level of 5%, how large \bar{Y} must be for us to reject?

Rejecting (or Not) a Null Hypothesis

- We will reject if the p-value implied by that value of \bar{Y} is smaller than 5%.
- In these one-sided tests, as we saw in slide 52, the formula for the p-value is:

$$1 - \Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}}\right)$$

- Therefore, we will reject H_0 if

$$1 - \Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}}\right) < 0.05$$

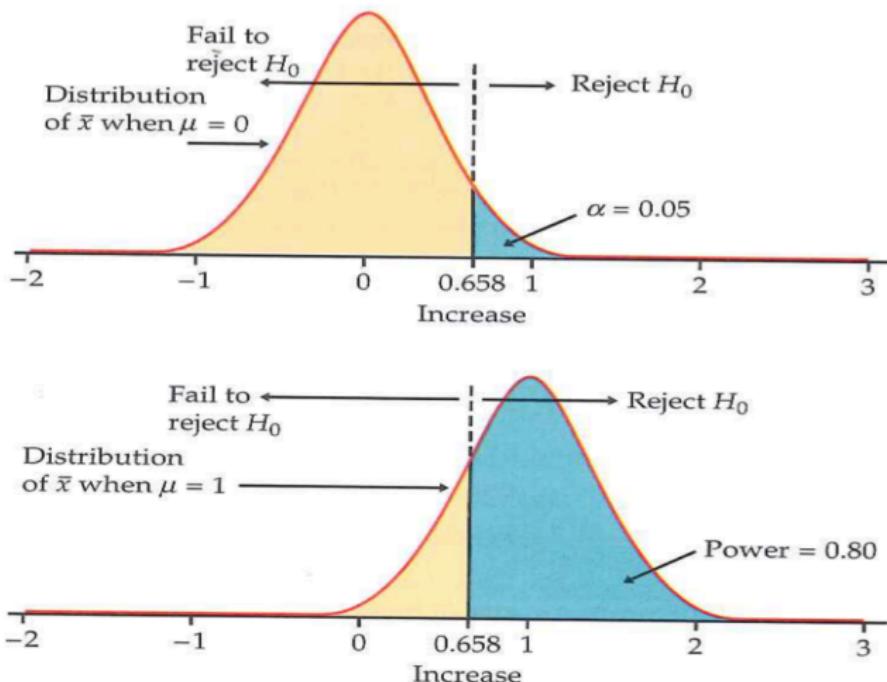
$$-\Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}}\right) < -0.95$$

$$\Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}}\right) > 0.95$$

$$\Phi\left(\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}}\right) > \Phi(1.645)$$

$$\frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}} > 1.645 \longrightarrow \bar{Y}^{obs} > 1.645 \frac{2}{\sqrt{25}} + 0 = 0.658.$$

Rejecting (or Not) a Null Hypothesis



The area to the right of 0.658 is the **rejection region**.

Rejecting (or Not) a Null Hypothesis

- Up until now, we have said that, given an observed random sample, we reject a null hypothesis if the p-value attached to that hypothesis in our observed sample is smaller than the significance level:

$$\text{if p-value} < \alpha, \text{ then reject } H_0.$$

- This rule is true for any kind of test about the population mean. The only thing that will change depending on whether we are doing a one-sided or a two-sided test is how we compute the p-value.
- However, for the particular example in slides 59 to 61, the mathematical derivation showed that we can rewrite the rejection rule as:

$$\text{if } \frac{\bar{Y}^{obs} - \mu_0}{\sigma_Y / \sqrt{n}} > 1.645, \text{ then reject } H_0,$$

where 1.645 was precisely the number such that $\Phi(1.645)$ equals our desired significance level, 0.95. The number 1.645 is said to be the **critical value** in a one-sided test associated to $\alpha = 0.95$.

Rejecting (or Not) a Null Hypothesis

- Example.
- Imagine that you want to test:

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu \neq 1.$$

Imagine also that your desired significance level is 0.05. You also know that the variance of Y in the population is equal to 4. If you observe a random sample of size 25, which values of \bar{Y} will make you reject H_0 in favor of H_1 ?

- We reject if p-value < 0.05, and, given that we are doing a two sided test, this expression becomes

$$2 \times \left(1 - \Phi \left(\frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}} \right) \right) < 0.05,$$

and plugging the specific values of our example,

$$2 \times \left(1 - \Phi \left(\frac{|\bar{Y}|}{2 / \sqrt{25}} \right) \right) < 0.05.$$

Rejecting (or Not) a Null Hypothesis

- Finding all the possible values of \bar{Y} that satisfy the previous inequality is annoying. Luckily, we can rewrite this inequality in a way that is easier to define which values of \bar{Y} satisfy it.
- From

$$2 \times \left(1 - \Phi \left(\frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}} \right) \right) < 0.05,$$

we can rewrite

$$\left(1 - \Phi \left(\frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}} \right) \right) < \frac{0.05}{2},$$

$$1 - \frac{0.05}{2} < \Phi \left(\frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}} \right)$$

$$97.5 < \Phi \left(\frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}} \right)$$

$$\Phi(1.96) < \Phi \left(\frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}} \right)$$

Rejecting (or Not) a Null Hypothesis

- and, therefore, we can rewrite the rejection rule as:

$$1.96 < \frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}},$$

where 1.96 is the **critical value** of this two-sided test. Using this expression, note that we will reject the null hypothesis if

$$\bar{Y} - \mu_0 > 1.96 \frac{\sigma_Y}{\sqrt{n}} \quad \rightarrow \quad \bar{Y} > \mu_0 + 1.96 \frac{\sigma_Y}{\sqrt{n}},$$

or

$$\bar{Y} - \mu_0 < -1.96 \frac{\sigma_Y}{\sqrt{n}} \quad \rightarrow \quad \bar{Y} < \mu_0 - 1.96 \frac{\sigma_Y}{\sqrt{n}}.$$

- Therefore, we reject the null hypothesis if

$$\bar{Y} \in \left\{ \left(-\infty, \mu_0 - 1.96 \frac{\sigma_Y}{\sqrt{n}} \right) \cup \left(\mu_0 + 1.96 \frac{\sigma_Y}{\sqrt{n}}, \infty \right) \right\},$$

where the term in brackets is the **rejection region**.

Rejecting (or Not) a Null Hypothesis

- After having seen a few examples, let's summarize the most important things you need to know in a systematic way.
- Imagine that you want to test

$$H_0 : \mu_Y = \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_0;$$

in a population in which $\mathbb{V}[Y] = \sigma_Y^2$, using a random sample whose average is \bar{Y}^{obs} , and with a desired significance level α .

- Then reject if

$$P\left(|Z| \geq \left|\frac{\bar{Y}^{\text{obs}} - \mu_0}{\sigma_Y / \sqrt{n}}\right|\right) < \alpha,$$

or, equivalently, reject if

$$\left|\frac{\bar{Y}^{\text{obs}} - \mu_0}{\sigma_Y / \sqrt{n}}\right| \geq Z_{\alpha/2}.$$

where $Z_{\alpha/2}$ is such that $\Phi(Z_{\alpha/2}) = 1 - (\alpha/2)$.

Rejecting (or Not) a Null Hypothesis

- Imagine that you want to test

$$H_0 : \mu_Y \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y > \mu_0;$$

in a population in which $\mathbb{V}[Y] = \sigma_Y^2$, using a random sample whose average is \bar{Y}^{obs} , and with a desired significance level α .

- Then reject if

$$P\left(Z \geq \frac{\bar{Y}^{\text{obs}} - \mu_0}{\sigma_Y / \sqrt{n}}\right) < \alpha,$$

or, equivalently, reject if

$$\frac{\bar{Y}^{\text{obs}} - \mu_0}{\sigma_Y / \sqrt{n}} \geq Z_\alpha,$$

where Z_α is such that $\Phi(Z_\alpha) = 1 - \alpha$.

Rejecting (or Not) a Null Hypothesis

- Imagine that you want to test

$$H_0 : \mu_Y \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y < \mu_0;$$

in a population in which $\mathbb{V}[Y] = \sigma_Y^2$, using a random sample whose average is \bar{Y}^{obs} , and with a desired significance level α .

- Then reject if

$$P\left(Z \leq \frac{\bar{Y}^{\text{obs}} - \mu_0}{\sigma_Y / \sqrt{n}}\right) < \alpha,$$

or, equivalently, reject if

$$\frac{\bar{Y}^{\text{obs}} - \mu_0}{\sigma_Y / \sqrt{n}} \leq -Z_\alpha,$$

where Z_α is such that $\Phi(Z_\alpha) = 1 - \alpha$.

CONFIDENCE INTERVAL FOR THE POPULATION MEAN

Confidence Interval

- A confidence interval with confidence level $1 - \alpha$ for a parameter is an interval computed from sample data by a **method** that has probability $1 - \alpha$ of producing an interval containing the true value of the parameter.
- Therefore, a confidence interval for the population mean is an interval $[\theta_1(Y_1, \dots, Y_n), \theta_2(Y_1, \dots, Y_n)]$ such that

$$Pr\left(\theta_1(Y_1, \dots, Y_n) \leq \mu \leq \theta_2(Y_1, \dots, Y_n)\right) = 1 - \alpha.$$

- In a particular realization of a random sample, the function $\theta_1(Y_1, \dots, Y_n)$ and $\theta_2(Y_1, \dots, Y_n)$ will become particular numbers $\hat{\theta}_1$ and $\hat{\theta}_2$. In this case, we will report our confidence interval as $[\hat{\theta}_1, \hat{\theta}_2]$.

Confidence Interval

- It is common to think that

$$Pr(\hat{\theta}_1 \leq \mu \leq \hat{\theta}_2) = 1 - \alpha,$$

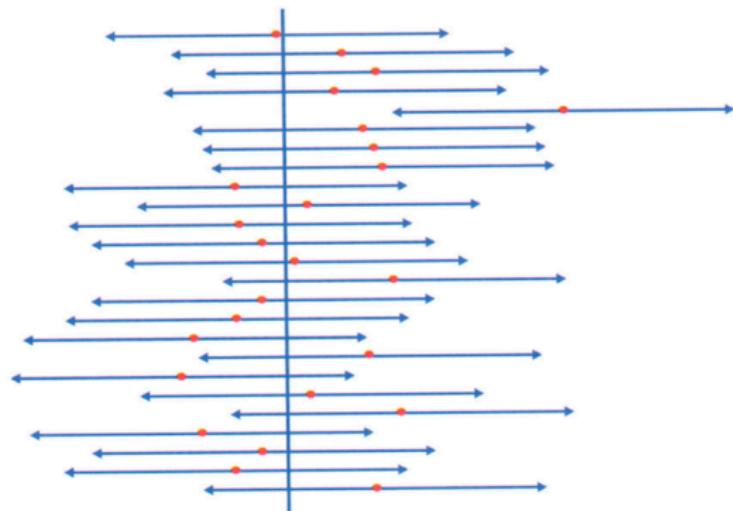
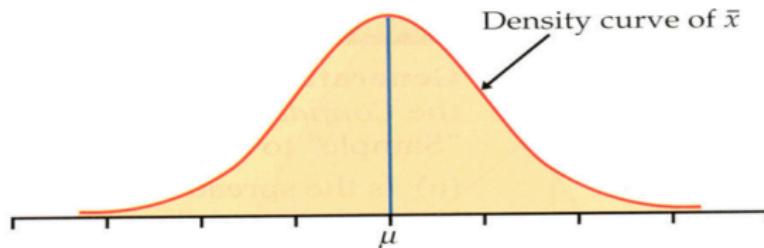
with $\hat{\theta}_1$ and $\hat{\theta}_2$ the particular realized values of the limits of the confidence interval for the population mean.

- However, the expression

$$Pr(\hat{\theta}_1 \leq \mu \leq \hat{\theta}_2) = 1 - \alpha,$$

does not make sense. If $\hat{\theta}_1$ and $\hat{\theta}_2$ are the realized values, then they are just numbers (not random variables). Also, μ is a number (not a random variable). Therefore, this probability is not defined over any random variable.

Confidence Interval



Confidence Interval for the Population Mean

- If the sample is large, a confidence interval for the population mean with confidence level $1 - \alpha$ may be constructed as:

$$[\bar{Y} - Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}}, \bar{Y} + Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}}]$$

where $Z_{\alpha/2}$ is a number such that $\Phi(Z_{\alpha/2}) = 1 - (\alpha/2)$.

- Standard confidence levels used in practice are:

$$1 - \alpha = 0.99 \quad \rightarrow \quad Z_{\alpha/2} = 2.57$$

$$1 - \alpha = 0.95 \quad \rightarrow \quad Z_{\alpha/2} = 1.96$$

$$1 - \alpha = 0.90 \quad \rightarrow \quad Z_{\alpha/2} = 1.65$$

- In case that the variance of Y is unknown, we can substitute σ_Y by $\hat{\sigma}_Y$ and, in large samples, the properties of the confidence interval described above remain invariant to the change.

Relationship between Two-Sided Tests and Conf. Interval

- Given confidence level α , imagine that you want to perform the test

$$H_0 : \mu_Y = \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_0.$$

- In slide 65, we saw that we reject the null if the observed sample average, \bar{Y} , is such that

$$Z_{\alpha/2} < \frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}}.$$

- Therefore, we do **not** reject the null if \bar{Y} is such that

$$Z_{\alpha/2} > \frac{|\bar{Y} - \mu_0|}{\sigma_Y / \sqrt{n}}.$$

Relationship between Two-Sided Tests and Conf. Interval

- Note that we can reorganize the last expression as

$$\bar{Y} - \mu_0 < Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}} \quad \rightarrow \quad \mu_0 > \bar{Y} - Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}},$$

and

$$\bar{Y} - \mu_0 > -Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}} \quad \rightarrow \quad \mu_0 < \bar{Y} + Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}}.$$

- Putting together these two expressions, we obtain exactly our confidence interval (see slide 73):

$$\bar{Y} - Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}} < \mu_0 < \bar{Y} + Z_{\alpha/2} \frac{\sigma_Y}{\sqrt{n}}.$$

- Conclusion: given a random sample, we reject the null $H_0 : \mu_Y = \mu_0$ in favor of the alternative hypothesis $H_1 : \mu_Y \neq \mu_0$ at the significance level α if and only if the confidence interval with confidence level $1 - \alpha$ does not include μ_0 .

TESTING HYPOTHESIS ABOUT THE DIFFERENCE OF MEANS

Testing Hypothesis about Difference of Means

- Before the 2006 Michigan gubernatorial primary, three political scientists isolated a group of potential voters and mailed them copies of their voting histories, listing the elections in which they participated and those they missed. Included were their neighbors' voting histories, too, along with a warning: after the polls closed, everyone would get an updated set.
- The three political scientists were interested in testing the null hypothesis that the probability to vote was the same for the voters that got their message in the mail as for the voters that did not get their message. Specifically, they wanted to test

$$H_0 : \mu_Y = \mu_N \quad \text{vs.} \quad H_1 : \mu_Y \neq \mu_N,$$

where μ_Y is the mean among the group of voters receiving the message of a random variable taking value 1 if an individual voted. The parameter μ_N denotes the mean of the same random variable among the group of voters who did not receive the message.

Testing Hypothesis about Difference of Means

- Imagine that you have access to data on a random sample of potential voters who got the message and a random sample of potential voters who did not get the message.
- Specifically, imagine that you have data on n_Y potential voters who got the message and n_N voters who did not get the message. Both n_Y and n_N are “large”.
- How would you test the null hypothesis?
- As in the case in which we were testing the null hypothesis, we need an estimator whose distribution is known. The CLT guarantees that the distribution of the sample mean is asymptotically normal and, therefore:

$$\bar{X}_Y \sim \mathbb{N}\left(\mu_Y, \frac{\sigma_Y^2}{n_Y}\right) \quad \text{and} \quad \bar{X}_N \sim \mathbb{N}\left(\mu_N, \frac{\sigma_N^2}{n_N}\right),$$

where \bar{X}_Y is the sample average among the set of potential voters who got the mail message (and similarly for \bar{X}_N among those who did not get the message).

Testing Hypothesis about Difference of Means

- Any observation in a random sample is independent of any observation from a different random sample (be sure you understand this).
- Therefore, any function of the observations in one random sample is independent of any function of the observations of a different random sample.
- Specifically, \bar{X}_Y and \bar{X}_N are independent.
- Therefore,

$$\bar{X}_Y - \bar{X}_N \sim \mathbb{N}\left(\mu_Y - \mu_N, \frac{\sigma_Y^2}{n_Y} + \frac{\sigma_N^2}{n_N}\right).$$

- Similarly, in large samples,

$$\bar{X}_Y - \bar{X}_N \sim \mathbb{N}\left(\mu_Y - \mu_N, \frac{\hat{\sigma}_Y^2}{n_Y} + \frac{\hat{\sigma}_N^2}{n_N}\right).$$

- Using this distribution, we can both test the null hypothesis $H_0 : \mu_Y = \mu_N$ as well as build confidence intervals for $\mu_Y - \mu_N$.

Testing Hypothesis about Difference of Means

- Note that we can rewrite

$$H_0 : \mu_Y = \mu_N,$$

as

$$H_0 : \mu_Y - \mu_N = 0.$$

- Therefore, we can apply the same logic we learned when testing the null hypothesis $H_0 : \mu = \mu_0$.
- Specifically, we reject the null hypothesis if

$$2 \times \left(1 - \Phi \left(\frac{|\bar{X}_Y - \bar{X}_N - 0|}{\sqrt{\frac{\hat{\sigma}_Y^2}{n_Y} + \frac{\hat{\sigma}_N^2}{n_N}}} \right) \right) < 0.05.$$

Testing Hypothesis about Difference of Means

- More generally, if we were testing

$$H_0 : \mu_Y - \mu_N = \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y - \mu_N = \mu_0,$$

we would reject the null hypothesis if

$$2 \times \left(1 - \Phi \left(\frac{|\bar{X}_Y - \bar{X}_N - \mu_0|}{\sqrt{\frac{\hat{\sigma}_Y^2}{n_Y} + \frac{\hat{\sigma}_N^2}{n_N}}} \right) \right) < 0.05.$$

- A $1 - \alpha$ confidence interval for $\mu_Y - \mu_N$ is

$$\left[\bar{X}_Y - \bar{X}_N - Z_{\alpha/2} \sqrt{\frac{\hat{\sigma}_Y^2}{n_Y} + \frac{\hat{\sigma}_N^2}{n_N}}, \bar{X}_Y - \bar{X}_N + Z_{\alpha/2} \sqrt{\frac{\hat{\sigma}_Y^2}{n_Y} + \frac{\hat{\sigma}_N^2}{n_N}} \right],$$

where $Z_{\alpha/2}$ is the number such that $1 - \Phi(Z_{\alpha/2}) = \alpha/2$.

- How would you test

$$H_0 : \mu_Y - \mu_N \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu_Y - \mu_N < \mu_0 ?$$

WWS 507c: Quantitative Analysis

Lecture 11: Causal Effect and Treatment Variables +
Regression with One Covariate

Princeton University

November 11, 2014

CONNECTING DIFFERENCES-OF-MEANS ESTIMATION TO ESTIMATION OF CAUSAL EFFECTS

Introduction

- Imagine you are working for the Food and Drug Administration (FDA).
- The FDA is dealing with a new malaria vaccine that Bayer has manufactured.
- The World Malaria Report of the WHO shows that in China, 40% to 60% of the population are at risk of suffering malaria.
- Bayer collected a **random sample** of size 300 from the Chinese population, assigned the new vaccine to a subsample of size 150, and a placebo to the rest of the sample. The results are contained in the following table:

	vaccine	placebo
N	150	150
M	25	75

where N denotes the number of individuals that got each treatment, and M denotes the number of individuals that suffered malaria.

Testing a Difference in Means

- Your boss at the FDA has asked you to determine whether the vaccine has a positive impact on reducing the incidence of malaria.
- You can formulate the task your boss has assigned to you as testing

$$H_0 : \mu_V - \mu_P \geq 0, \quad \text{vs.} \quad H_1 : \mu_V - \mu_P < 0,$$

where μ_V is the share of individuals suffering malaria in a population of vaccinated individuals, and μ_P is the corresponding share for a population of non-vaccinated individuals.

- In particular, both μ_V and μ_P are means of a Bernoulli random variable, Y_i , that takes value 1 if individual i suffers malaria, and value 0 otherwise. Using conditional expectations, we can rewrite them as:

$$\mathbb{E}[Y_i | V_i = 1] = \mu_V \quad \text{and} \quad \mathbb{E}[Y_i | V_i = 0] = \mu_P,$$

where V_i is a dummy variable taking value 1 if i has received the vaccine.

Testing a Difference in Means

- From the CLT, we know that,

$$\bar{Y}_V - \bar{Y}_P \approx \mathbb{N}(\mu_V - \mu_P, \frac{\sigma_V^2}{N_V} + \frac{\sigma_P^2}{N_P})$$

and

$$\frac{(\bar{Y}_V - \bar{Y}_P) - (\mu_V - \mu_P)}{\sqrt{\frac{\sigma_V^2}{N_V} + \frac{\sigma_P^2}{N_P}}} \approx \mathbb{N}(0, 1).$$

- Given that Y_i is a binary random variable

$$\sigma_V^2 = \mu_V(1 - \mu_V) \quad \text{and} \quad \sigma_P^2 = \mu_P(1 - \mu_P).$$

- Therefore, **under the null hypothesis** $H_0 : \mu_V - \mu_P = 0$, it holds that

$$\frac{\bar{Y}_V - \bar{Y}_P}{\sqrt{\mu(1 - \mu) \left(\frac{1}{N_V} + \frac{1}{N_P} \right)}} \approx \mathbb{N}(0, 1).$$

where $\mu = \mu_V = \mu_P$.

Testing a Difference in Means

- The parameter μ is unknown. In order to do a test of the null hypothesis $H_0 : \mu_V - \mu_P = 0$, we need to find a consistent estimate of μ . We can use

$$\bar{Y}^{obs} = \frac{150}{300} \frac{25}{150} + \frac{150}{300} \frac{75}{150} = \frac{1}{3}.$$

- Therefore, under the null hypothesis,

$$\frac{\bar{Y}_V - \bar{Y}_P}{\sqrt{\frac{2}{6}(1 - \frac{2}{6})\left(\frac{1}{150} + \frac{1}{150}\right)}} = \frac{\bar{Y}_V - \bar{Y}_P}{0.0544} \approx \mathbb{N}(0, 1)$$

- The p-value from the test is

$$P(Z < \frac{\bar{Y}_V^{obs} - \bar{Y}_P^{obs}}{0.0544}) = P(Z < \frac{-0.33}{0.0544}) \approx 0.$$

Testing a Difference in Means

- Therefore, for any generally used significance level (e.g. $\alpha = 0.01$), we can reject the null hypothesis that the population infection rate in the vaccinated population is the same as the infection rate in the non-vaccinated population, in favor of the alternative hypothesis that the malaria infection rate is lower in the vaccinated population.
- This test shows a very strong negative **association** or **correlation** between a dummy variable indicating whether an individual has received the new vaccine and a dummy variable indicating whether this individual suffered malaria.
- Using this evidence, can you conclude that the new vaccine manufactured by Bayer is effective at reducing malaria and, therefore, the FDA should grant their approval?

NO!

NO!

NO!

NO!

NO!

NO!

NO!

NO!

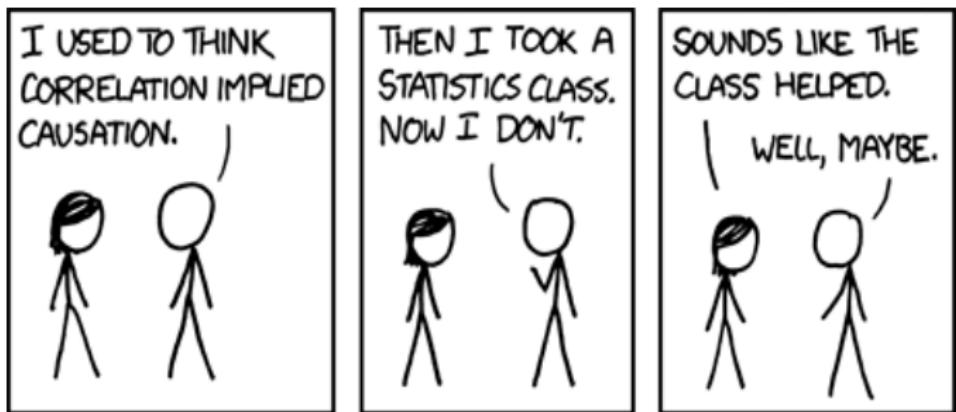
NO!

and

NO!!!!!!

Interpreting a Difference in Means

- The difference in the population means, μ_V and μ_P , does not necessarily reflect a **causal** relationship between a dummy variable indicating whether an individual has received the new vaccine and a dummy variable indicating whether this individual suffered malaria. It just reflects a **correlation** between those two variables.



Interpreting a Difference in Means

- Why the empirical evidence showing that the infection rate is lower in the subpopulation that took the vaccine is not enough evidence of a causal impact of the vaccine on the probability of suffering malaria?
- There are other possible explanations for this difference in infection rates:
 - ① If Bayer made access to the new vaccine voluntarily, it is likely that those individuals that are more concerned with malaria will be those opting into the vaccine. However, it is likely that these individuals are also more likely to use mosquito-nests, apply an insect repellent on their skin and clothes, consult a physician if they get sick, etc. And it could be that their lower infection rate is due to these other preventive measures (and not to the new vaccine).
 - ② If Bayer selected who got the vaccine and who the placebo, it is possible that Bayer decided to give the vaccine to those patients living in areas with relatively low incidence of malaria. Again, these individuals would have lower infection rates, independently of the vaccine.

Causal Effect: Terminology

- How can we **empirically** learn about the **causal effect** of some **treatment variable** X (e.g. receiving a malaria vaccine) on some **outcome variable** Y (e.g. suffering malaria)? We need variation in Y that is only due to X .
- Ideally, we would like to see the same individuals with and without treatment (e.g. with and without vaccine), and observe whether their outcome variable varies with the treatment (e.g. whether they are more likely suffer malaria in one case than in another). In this case, it would be easy to empirically determine the effect of the treatment (e.g. the effect of the vaccine).
- Given that we cannot observe the same individual with and without vaccine, the empirical evidence supporting the **causal effect** of the vaccine has to be based on inter-personal comparisons: are the individuals who received the vaccine less likely to suffer malaria than those who did not receive it?
- Inter-personal comparisons are informative about the effect of a treatment on an outcome if individuals in **treatment** and **control groups** are similar in any other dimension that affects the outcome variable.

Defining Causal Effects: Random Experiments

- How do we obtain variation in an outcome variable that is exclusively due to a treatment variable (e.g. variation in the probability of suffering malaria exclusively due to malaria vaccine)? \leftrightarrow How do we obtain treatment and control groups that are identical except in their treatment level?
- The ideal way to obtain this variation is through a **randomized controlled experiment**. In a randomized controlled experiment, the treatment is assigned randomly. This eliminates the possibility of a systematic relationship between, for example, the likelihood an individual uses mosquito-nests and the likelihood she gets a vaccine.
- The concept of an ideal randomized controlled experiment is useful because it gives a **definition** of a causal effect.
- In the particular case of a **binary treatment variable**, the **causal effect** on an outcome variable is **defined** as the difference in population means between a treatment and a control group (e.g. $\mu_V - \mu_P$) that **would** arise if the assignment of individuals to treatment and controls groups **was** random.

Defining Causal Effects: Random Experiments

- The difference in population means between a treatment and control group (e.g. $\mu_V - \mu_P$) reflects the causal effect of a binary treatment **if** the assignment of individuals to treatment and control groups is random.
- Using data from a sample and the tools of hypothesis testing to reject the null hypothesis $H_0 : \mu_V - \mu_P = 0$ implies empirical evidence in favor of the beneficial effects of a new vaccine **if**: (a) the sample is random; and, (b) the assignment of individuals to treatment and control groups is random.
- Note that the statements in the two previous bullet points are *if* conditions (not *if and only if* conditions). The reason is that, while the concept of an ideal randomized controlled experiment is useful because it gives a **definition** of a causal effect, this does not mean that we need to have data coming from randomized controlled experiments in order to estimate causal effects. We will learn how to **estimate** causal effects in cases in which the sample or data available is not coming from a randomized controlled experiment.

Connecting Differences-of-Means Estimation to Regression

- We have seen above that the causal effect of a **binary treatment variable** X on an outcome variable Y can be **defined** as the difference in the mean of Y between the treatment group ($X = 1$) and the control group ($X = 0$), **when the assignment of individuals to treatment and control groups is random**. Mathematically:

$$\text{causal effect of } X \text{ on } Y = \mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0],$$

if the assignment of values of X to individuals is random. Therefore, the causal effect of a **binary treatment variable** is defined as the difference in the expectation of the outcome variable conditional on the treatment variable **in an ideal randomized controlled experiment.**

- If the assignment of the treatment variable X is not random, then both $\mathbb{E}[Y|X = 1]$ and $\mathbb{E}[Y|X = 0]$ are still well-defined objects, but their difference is not the causal effect of X on Y .

Connecting Differences-of-Means Estimation to Regression

- The difference in sample means $\bar{Y}_1 - \bar{Y}_0$, with

$$\bar{Y}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i},$$
$$\bar{Y}_0 = \frac{\sum_{i=1}^n (1 - X_i) Y_i}{\sum_{i=1}^n (1 - X_i)},$$

is an unbiased, consistent, and efficient estimator of the causal effect of X on Y as long as: (a) $\{(X_i, Y_i), i = 1, \dots, n\}$ are *i.i.d*; (b) the assignment of the value of X_i to each individual i in the sample is random.

- If the assignment of X_i to i is not random, then $\bar{Y}_1 - \bar{Y}_0$ is still an unbiased, consistent, and efficient estimator of the difference in means,

$$\mathbb{E}[Y|X = 1] - \mathbb{E}[Y|X = 0],$$

but this difference in means no longer reflects the causal effect of X on Y .

Connecting Differences-of-Means Estimation to Regression

- In many cases, we care about the causal effect of variables that are not binary.
 - As an example, imagine that you are working for the U.S. Department of Education and the Secretary of Education asks you for advice on the ideal class size. In particular, the Secretary asks you to evaluate the causal effect of class size on the grade obtained by fifth graders in a standardized test.
 - In order to answer this question, assume you only have data on class size and test scores for a random sample of fifth graders.
 - The treatment variable of interest, X , is class size. The outcome variable of interest, Y , is test grades. In your sample, the variable X takes values between 15 and 30, and the variable Y takes values between 600 and 700.
 - For any two numbers, x_1 and x_2 , one can define the treatment or causal effect of interest as the effect on grades of switching class sizes from x_1 to x_2 .
- If the allocation of students to classes of different size is random**, then we can define this causal effect as

$$\mathbb{E}[Y|X = x_1] - \mathbb{E}[Y|X = x_2].$$

Connecting Differences-of-Means Estimation to Regression

- If we observe a random sample of grades for students that have been randomly assigned to classes of different size, then we can estimate the causal effect of switching a student from a class of any given size x_1 to a class of any other size x_2 . For a given pair of sizes (x_1, x_2) , the difference in sample means

$$\bar{Y}_1 - \bar{Y}_2,$$

with

$$\bar{Y}_1 = \frac{\sum_{i=1}^n \mathbb{1}\{X_i = x_1\} Y_i}{\sum_{i=1}^n \mathbb{1}\{X_i = x_1\}},$$
$$\bar{Y}_2 = \frac{\sum_{i=1}^n \mathbb{1}\{X_i = x_2\} Y_i}{\sum_{i=1}^n \mathbb{1}\{X_i = x_2\}},$$

is an estimator of the causal effect $\mathbb{E}[Y|X = x_1] - \mathbb{E}[Y|X = x_2]$.

- This approach requires computing as many treatment effects as possible pairs (x_1, x_2) there exist.

CONSTANT CAUSAL EFFECTS

Constant Causal Effects

- When the variable X can take a large finite set of values, the approach described in the previous slide is feasible but it requires estimating a large number of difference-in-means estimators.
- When the variable X takes an infinite number of possible values (e.g. X is continuous), the approach described in the previous slides is impossible.
- How can we estimate then the causal effect of increasing a continuous random variable X on a continuous random variable Y ?
- In this lecture, we will focus on answering this question for the particular case in which the causal effect on Y of increasing the random variable X in one unit is constant: β_1 .
- In other words, in this lecture, we will focus on the case in which there is an unknown number β_1 such that the causal effect on grades of increasing class sizes from x_1 to $x_1 + 1$ is equal to β_1 , for any possible value of x_1 .

Constant Causal Effects

- As we indicated above, one way to think about causal effects is through the concept of randomization.
- In this lecture, we will focus on examples in which, if a treatment variable X had been randomly assigned to the individuals in the population, then the difference in the expectation of a variable Y between those individuals with assigned values of X equal to x_1 and with assigned values of X equal to $x_1 + 1$ would be equal to a constant β_1 , independently of the particular value of x_1 .
- In this lecture, we will learn how to use a random sample to estimate the causal effect of a random variable X on an outcome variable Y only in the particular case in which the causal effect of increasing X in one unit is equal to a constant. We denote this constant as β_1 .
- In future lectures, we will learn how to estimate this causal effect in the more general case in which the causal effect of increasing X in one unit is different depending on whether we increase it from x_1 to $x_1 + 1$ or from x_2 to $x_2 + 1$, for two different numbers x_1 and x_2 .

Constant Causal Effects

- If the causal effect of X_i (e.g. class size) on Y_i (e.g. grades) is constant, we can write the relationship between Y_i and X_i as

$$Y_i = \beta_1 X_i + u_i,$$

where β_1 denotes the causal effect of increasing X_i in one unit on Y_i , and u_i captures all the other factors that could be affecting Y_i beyond X_i . Examples of these other factors could be: (a) weather conditions; (b) quality of teachers; (c) students' effort, etc.

- Denoting the expectation of u_i as β_0 , we can rewrite

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where, by definition, $\mathbb{E}[\varepsilon_i] = 0$. Note that this equation is true even if X_i has not been randomly assigned in the population.

Constant Causal Effects

- In the specific case in which the different values of X are randomly assigned to the different individuals i in the population (e.g. students are randomly assigned to classes of different sizes), it will be true that X_i and ε_i are independent. A consequence of X_i and ε_i being independent is that

$$\mathbb{E}[\varepsilon_i | X_i] = 0.$$

- Therefore, **if the different values of X are randomly assigned to the different individuals i in the population, then**

$$\mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i.$$

- The reason is that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{plus} \quad \mathbb{E}[\varepsilon_i | X_i] = 0,$$

implies

$$\mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i.$$

Constant Causal Effects

- Is it possible that the two random variables ε_i and X_i verify that

$$\mathbb{E}[\varepsilon_i | X_i] = 0,$$

even if X_i has not been randomly assigned in the population? **Yes!**

- Even if the government has not purposefully assigned students to class sizes through a randomization process, it is still perfectly possible that, for example, (a) school districts with better weather do not tend to have larger or smaller class sizes; (b) high quality professors do not tend to be assigned to larger or smaller class sizes; (c) students study the same independently of whether they belong to a large or a small class...
- Therefore, random assignment of X in the population is **sufficient but not necessary** for the relationships

$$\mathbb{E}[\varepsilon_i | X_i] = 0, \Leftrightarrow \mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i$$

to be true in the population of interest.

Constant Causal Effects

- If there is a variable other than X that affects the outcome variable Y beyond X (e.g. income per capita in a school district may affect school grades through channels other than class sizes, because kids of richer parents may be more likely to pay for private teachers that help kids that struggle in school) and this variable is correlated with X (e.g. income per capita in a school district may be correlated with smaller class sizes because higher property taxes help pay for more teachers), then

$$\mathbb{E}[\varepsilon_i | X_i] \neq 0, \Leftrightarrow \mathbb{E}[Y_i | X_i] \neq \beta_0 + \beta_1 X_i.$$

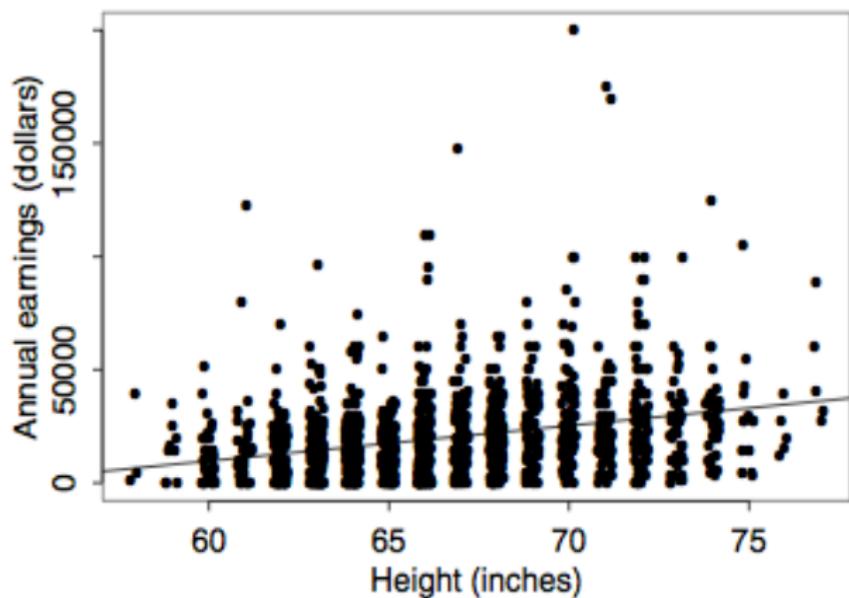
- Why is it important whether $\mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i$ or $\mathbb{E}[Y_i | X_i] \neq \beta_0 + \beta_1 X_i$?
- Because only in the case in which $\mathbb{E}[Y_i | X_i] = \beta_0 + \beta_1 X_i$ we can use a random sample of the vector (X_i, Y_i) from the population of interest (e.g. a random sample of school districts in the U.S.) to compute consistent estimates of the causal effect β_1 .

Constant Causal Effects

- If the variables included in ε_i are correlated with the treatment variable of interest X_i , then $\mathbb{E}[\varepsilon|X_i] \neq 0$.
- In this case, the conditional expectation of Y_i on X_i , $\mathbb{E}[Y_i|X_i]$, still exists and is well defined but it will not be equal to $\beta_0 + \beta_1 X_i$.
- If $\mathbb{E}[\varepsilon|X_i] \neq 0$, even if we have a random sample of the vector (X_i, Y_i) from the population of interest, we will **not** be able to obtain consistent estimates of the causal effect β_1 .
- General conclusion: given a random sample, we can always obtain consistent estimates of $\mathbb{E}[Y_i|X_i]$. However, only in the case in which $\mathbb{E}[Y_i|X_i]$ is a known function of the causal effect of interest β_1 , having consistent estimates of $\mathbb{E}[Y_i|X_i]$ will be useful to learn something about β_1 .

Example

- In this example, does the conditional expectation reflect the causal effect of height on annual income?



ESTIMATING CONSTANT CAUSAL EFFECTS: OLS ESTIMATOR

Finding a Good Estimator

- For the rest of this lecture, assume that:

- the causal effect of a variable X (e.g. number of years of education) on a variable Y (e.g. annual wages) is given by

$$\beta_0 + \beta_1 X_i.$$

- All factors affecting wages besides education are mean independent of education:

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i.$$

- Both X and Y have finite fourth moments

$$0 < \mathbb{E}(X_i^4) < \infty \quad 0 < \mathbb{E}(Y_i^4) < \infty.$$

- We observe a random sample (X_i, Y_i) of size n from the population of interest:

$$\{(X_i, Y_i); i = 1, \dots, n\} \text{ are iid.}$$

- These 4 assumptions are the so-called **OLS assumptions**.

Finding a Good Estimator

- Given these assumptions, there is one estimator of (β_0, β_1) that is called the OLS estimator and that will be
 - unbiased,

$$\mathbb{E}[\hat{\beta}_0] = \beta_0$$

$$\mathbb{E}[\hat{\beta}_1] = \beta_1$$

- its variance decreases in the sample size n ,
- consistent,

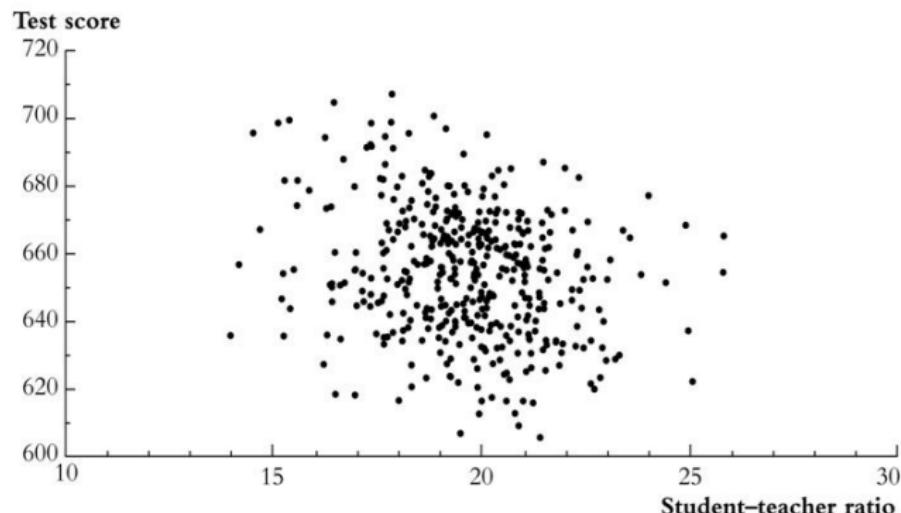
$$\hat{\beta}_0 \xrightarrow{n \rightarrow \infty} \beta_0$$

$$\hat{\beta}_1 \xrightarrow{n \rightarrow \infty} \beta_1$$

- asymptotically normal; this is important because it allows us to use the OLS estimators $(\hat{\beta}_0, \hat{\beta}_1)$ to perform test of hypothesis about (β_0, β_1) and to build confidence intervals.

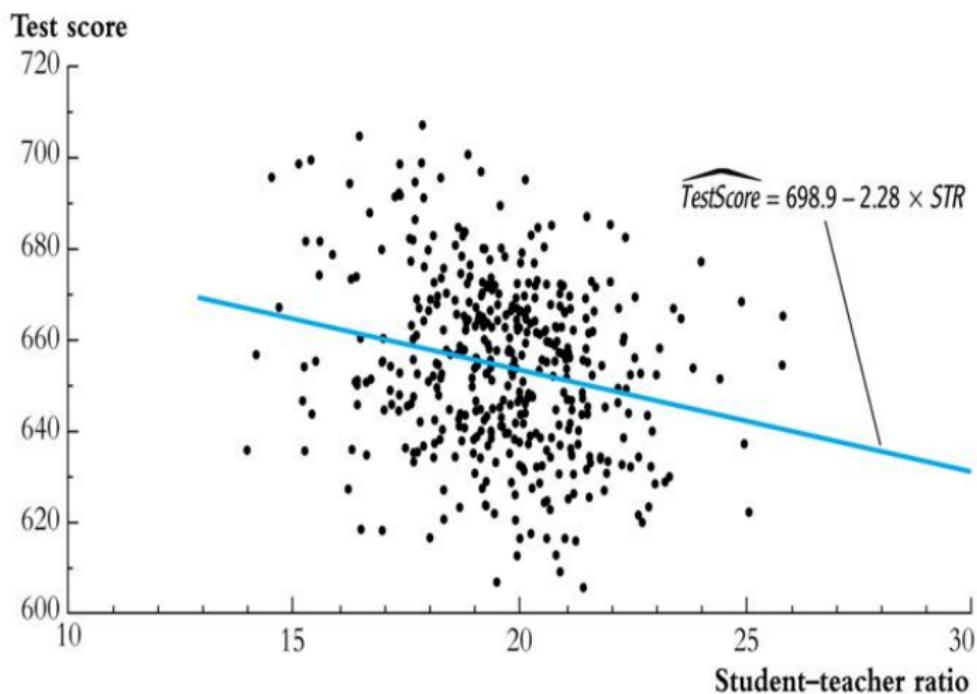
OLS Estimator

- Given a random sample of (X_i, Y_i) from the population of interest, what does the OLS estimator of (β_0, β_1) do?
- In very informal terms, it just computes the straight line that minimizes the sum of the squared vertical distances from each observation to such line. If the sample is...



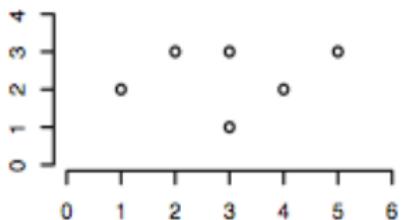
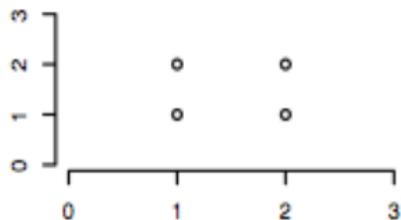
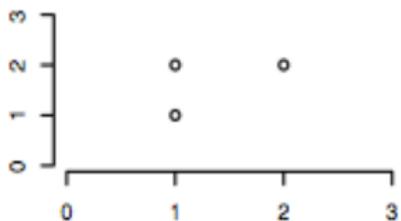
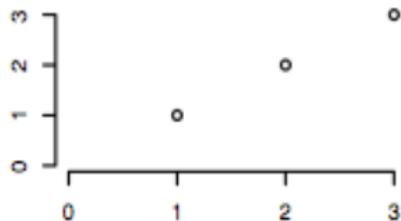
OLS Estimator

- ...then the regression line is...



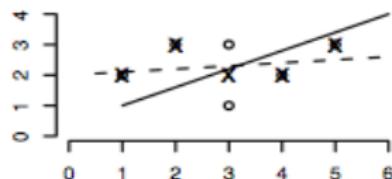
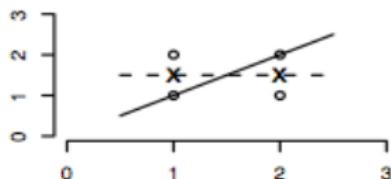
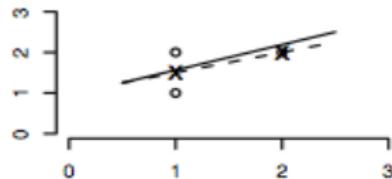
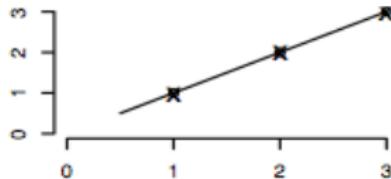
OLS Estimator

- Imagine these are your samples, how do you think the regression line will look like?



OLS Estimator

- Were you right?



[the right regression line is the dotted one; the solid line is a potential guess]

OLS Estimator

- Given a sample $\{(Y_i, X_i), i = 1 \dots, n\}$, we **define** the OLS estimator as:

$$\hat{\beta}_1 = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$$

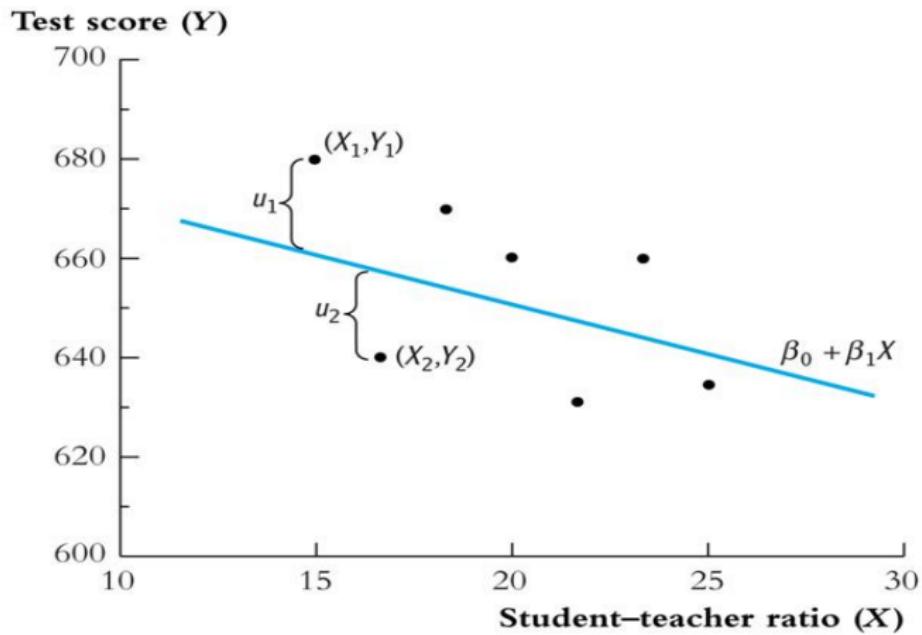
$$\hat{\beta}_0 = \frac{1}{n} \sum_{i=1}^n Y_i - \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n X_i = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- The OLS estimator, $(\hat{\beta}_0, \hat{\beta}_1)$, chooses the regression coefficients, (b_0, b_1) , so that the estimated regression line minimizes the sum of the squared mistakes made in predicting Y given X

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

- The OLS predicted values are: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- The OLS residuals are: $Y_i - \hat{Y}_i$.

OLS Estimator



OLS Estimator

OLS regression: STATA output

```
regress testscr str, robust
```

Regression with robust standard errors

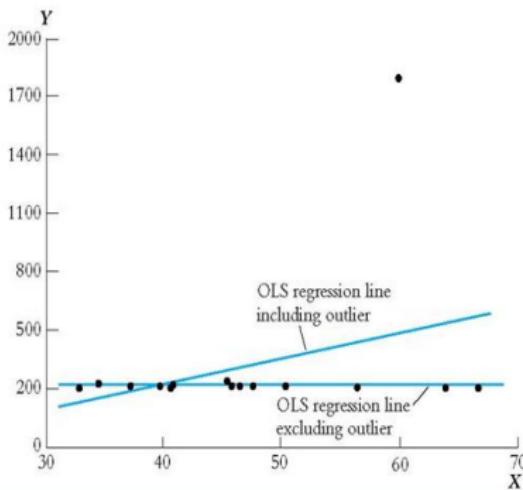
Number of obs = 420
F(1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581

	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

OLS Estimator: Sensitive to Outliers

FIGURE 4.5 The Sensitivity of OLS to Large Outliers

This hypothetical data set has one outlier. The OLS regression line estimated with the outlier shows a strong positive relationship between X and Y , but the OLS regression line estimated without the outlier shows no relationship.



Finite 4th moments imply that outliers are unlikely. Outliers may have a large impact on the OLS estimator of β_1 .

EXPECTATION AND VARIANCE OF OLS ESTIMATOR

Expectation and Variance of OLS Estimator

- Under the 4 OLS assumptions, the OLS estimator of (β_0, β_1) is unbiased

$$\mathbb{E}[\hat{\beta}_0] = \beta_0$$

$$\mathbb{E}[\hat{\beta}_1] = \beta_1$$

- Under the 4 OLS assumptions, the variance of the OLS estimator of β_1 is

$$\mathbb{V}(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \mathbb{E}\left[\frac{1}{n} \frac{\frac{1}{n} \sum_{i=1}^n \mathbb{V}[\varepsilon_i | X](X_i - \bar{X})^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]^2}\right],$$

where $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$.

- None of the 4 OLS assumptions imposes any restriction on $\mathbb{V}[\varepsilon_i | X]$. Under the additional restriction that

$$\mathbb{V}[\varepsilon_i | X] = \sigma^2,$$

the variance of the OLS estimator of β_1 is

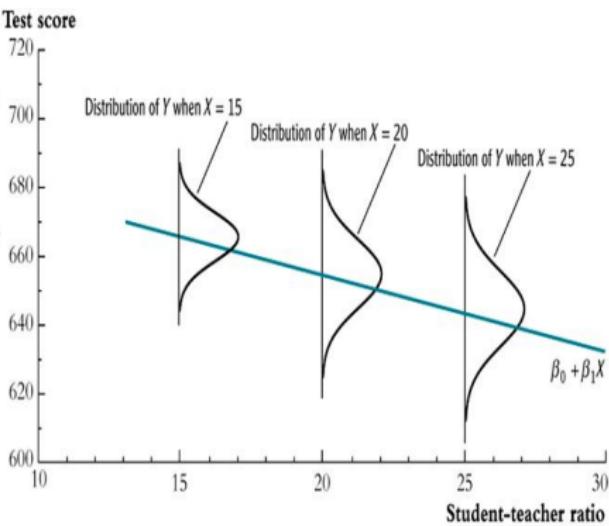
$$\sigma_{\hat{\beta}_1}^2 = \mathbb{E}\left[\frac{1}{n} \frac{\frac{1}{n} \sum_{i=1}^n \sigma^2 (X_i - \bar{X})^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]^2}\right] = \mathbb{E}\left[\frac{1}{n} \frac{\sigma^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}\right] = \mathbb{E}\left[\frac{1}{n} \frac{\sigma^2}{\hat{\sigma}_X^2}\right]$$

Heteroskedasticity

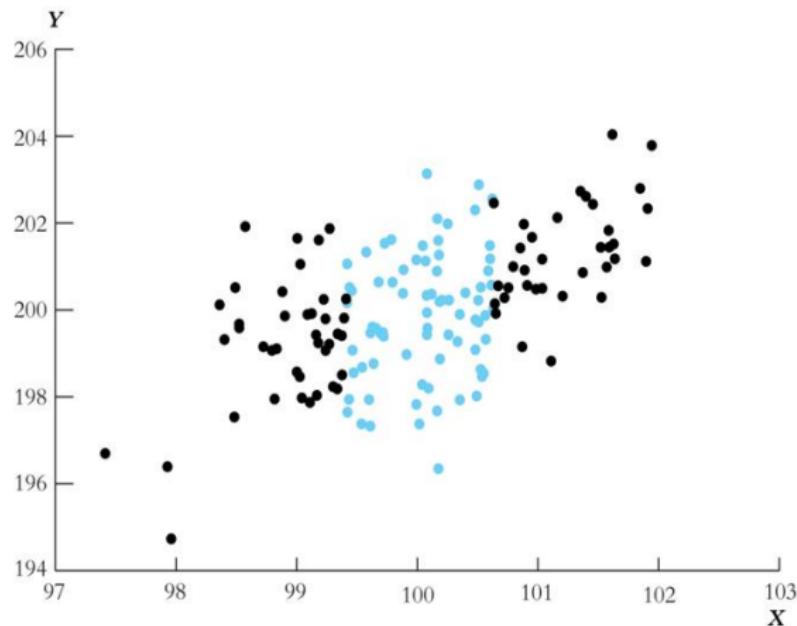
- Whenever the variance of ϵ_i does not depend on X_i , we say that the error term is **homoskedastic**. If the variance of ϵ_i depends on X_i , we say that the error term is **heteroskedastic**.

FIGURE 4.7 An Example of Heteroskedasticity

Like Figure 4.4, this shows the conditional distribution of test scores for three different class sizes. Unlike Figure 4.4, these distributions become more spread out (have a larger variance) for larger class sizes. Because the variance of the distribution of u given X , $\text{var}(u|X)$, depends on X , u is heteroskedastic.



Which Sample Do You Prefer?



Variance of X is larger for the black dots than for the blue ones: $\sigma_X^{2blue} < \sigma_X^{2black}$.

Asymptotic Variance of the OLS Estimator

- Independently of whether the error term is homoskedastic or heteroskedastic, the variance of the OLS estimator goes to 0 as the sample size goes to ∞ .
- It is easier to see this mathematically in the case in which the error is homoskedastic:

$$\sigma_{\hat{\beta}_1}^2 = \mathbb{E} \left[\frac{1}{n} \frac{\sigma^2}{\hat{\sigma}_X^2} \right],$$

and

$$\hat{\sigma}_X^2 \xrightarrow{n \rightarrow \infty} \sigma_X^2,$$

so

$$\sigma_{\hat{\beta}_1}^2 \xrightarrow{n \rightarrow \infty} \frac{1}{n} \frac{\sigma^2}{\sigma_X^2} \xrightarrow{n \rightarrow \infty} 0.$$

Estimates of the Variance of the OLS Estimator

- The variance of the OLS estimator, $\sigma_{\hat{\beta}_1}^2$, depends on elements that are usually unknown:
 - under heteroskedasticity: it depends on $\mathbb{V}[\varepsilon|X]$
 - under homoskedasticity: it depends on $\mathbb{V}[\varepsilon] = \sigma^2$
- Using a random sample $\{(X_i, Y_i); i = 1, \dots, n\}$, we can construct estimates of the variance of the OLS estimator.
 - under heteroskedasticity:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 (X_i - \bar{X})^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right]^2}$$

This estimator is known as the **robust** standard error or Eicker-Huber-White standard error.

- under homoskedasticity:

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Why You Should Always Use the Robust Standard Error?

- In large samples, the robust standard error converges to the true variance of the OLS estimator $\hat{\beta}_1$.
- Important: this is true independently of whether the errors are truly heteroskedastic or homoskedastic.
- Even if the errors are homoskedastic, the robust estimator of the variance of the OLS estimator $\hat{\beta}_1$ will converge to the true variance in large samples.
- On the contrary, the estimator of the standard error that assumes homoskedasticity will converge to the true variance only if the assumption of homoskedasticity is right.
- The only advantage of the estimator of $\sigma_{\hat{\beta}_1}^2$ that assumes homoskedasticity is that it is an efficient estimator in the specific case in which the errors are actually homoskedastic. This is not very important in large samples. Given that we will always deal with large samples, there is no reason to use the estimator of the variance of $\hat{\beta}_1$ that assumes homoskedasticity.

OLS Estimator

OLS regression: STATA output

```
regress testscr str, robust
```

Regression with robust standard errors

Number of obs = 420
F(1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581

	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

HYPOTHESIS TESTING AND CONFIDENCE INTERVALS

Distribution of OLS Estimator in Large Samples

- In large samples, the distribution of the OLS estimator is close to normal.
- Independently of whether the errors are homoskedastic or heteroskedastic,

$$\hat{\beta}_1 \approx \mathbb{N}\left(\beta_1, \hat{\sigma}_{\hat{\beta}_1}^2\right),$$

with

$$\hat{\sigma}_{\hat{\beta}_1}^2 = \frac{1}{n} \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 (X_i - \bar{X})^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]^2}.$$

- Therefore, independently of whether the errors are homoskedastic or heteroskedastic

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \approx \mathbb{N}(0, 1) \quad \text{with} \quad \hat{\sigma}_{\hat{\beta}_1} = \sqrt{\frac{1}{n} \frac{\frac{1}{n} \sum_{i=1}^n \varepsilon_i^2 (X_i - \bar{X})^2}{\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right]^2}}.$$

Hypothesis Tests and Confidence Intervals

- Compute the p-value for the following test:

$$H_0 : \beta_1 = \gamma \quad \text{vs.} \quad H_1 : \beta_1 \neq \gamma,$$

for some given value γ . Using the asymptotic distribution of the t-statistic, we obtain:

$$P(|Z| > \left| \frac{\hat{\beta}_1^{obs} - \gamma}{\hat{\sigma}_{\hat{\beta}_1}^{obs}} \right|) = 2 \times (1 - \Phi(\frac{\hat{\beta}_1^{obs} - \gamma}{\hat{\sigma}_{\hat{\beta}_1}^{obs}})).$$

and we reject at 5% significance level if

$$2 \times (1 - \Phi(\frac{\hat{\beta}_1^{obs} - \gamma}{\hat{\sigma}_{\hat{\beta}_1}^{obs}})) < 0.05.$$

Confidence Interval for Predicted Effects of Changing X

- Compute a 95% confidence interval for the effect on the expected value of Y of changing X by a given amount Δx . The population change is $\beta_1 \Delta x$. An estimate of the predicted change is $\hat{\beta}_1 \Delta x$.
- The 95% confidence interval for β_1 is:

$$[\hat{\beta}_1 - 1.96 \times \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + 1.96 \times \hat{\sigma}_{\hat{\beta}_1}],$$

where $\Phi(1.96) = 0.975$.

- The 95% confidence interval for $\beta_1 \Delta x$ is:

$$[\hat{\beta}_1 - 1.96 \times \hat{\sigma}_{\hat{\beta}_1} \times \Delta x, \hat{\beta}_1 + 1.96 \times \hat{\sigma}_{\hat{\beta}_1} \times \Delta x].$$

OLS Estimator

OLS regression: STATA output

```
regress testscr str, robust
```

Regression with robust standard errors

Number of obs = 420
F(1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581

	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

WWS 507c: Quantitative Analysis

Lecture 12: Causal Effect and Treatment Variables: Intuition

Princeton University

November 16, 2014

Introduction

- The Secretary of Education would like to know the **impact** (i.e. **causal effect**) of class size on grades for 10th graders in the US.
- You only observe average grades and average class sizes by school district.
- **Unit of observation:** school district.
- **Population of interest:** all school districts in the U.S. (13,500 districts).
- **Treatment variable:** average class size in 10th grade.
- **Outcome variable:** average grades in standardized test of 10th graders.
- **Objective of your study:** estimate the causal effect of average class size on average grades in the population of all school districts in the U.S.
- The causal effect of changing average class size from 20 to 21 students is the difference between the mean grade (in the population) in those districts with average class size 21 and the mean grade in those districts with average class size 20 **in a hypothetical world in which students are randomly assigned to different class sizes.**

Population

- The average grade in each school district (Y) is a **linear** function of:
 - average class size (variable 1);
 - average income per capita (variable 2);
 - number of rainy days during the year (variable 3).
- Specifically:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

with $\beta_0 = 500$, $\beta_1 = -1$, $\beta_2 = 7$, $\beta_3 = 0.1$, and each X_i denotes deviations with respect to its mean across all school districts in the U.S.

- These values of $(\beta_1, \beta_2, \beta_3)$ denote the causal effect of X_1 , X_2 , and X_3 on Y .
- In this example, the **causal effect of class size on grades is equal to -1**.
- This is the number that the Secretary of Education would like to know. This is the **parameter of interest**.
- If the Secretary of Education were to decrease class sizes in 1 student in every school district, then average grades in all districts would increase in 1 unit.

Population

- Note that, from the equation in slide 3, we can write:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i,$$

with $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_{1i} = \beta_2 X_{2i} + \beta_3 X_{3i}$.

- Imagine that you correctly assume that the causal effect of X_{1i} on Y_i is linear:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i,$$

but you do not know which are the true values of β_0 and β_1 .

- Imagine also that you have information on a **random sample** of school districts and that, for each district, you only observe Y_i and X_{1i} .
 - In the next lecture, you will also observe X_{2i} , but not today.
- **Key question for today:** What can you learn about the true value of β_1 given your random sample of (Y_i, X_{1i}) and the knowledge that

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i?$$

Two Different Worlds

- **Key answer for today:** Expressed in three different ways:
- It depends on the expectation of ε_i conditional on X_{1i} .
- It depends on the expectation of X_{2i} and X_{3i} conditional on X_{1i} .
- It depends on the relationship between, on one side, the avg. income pc (X_{2i}) and number of rainy days (X_{3i}) and, on the other side, the avg. class size (X_{1i}) **in the population**.
- We will consider two worlds:
 - World 1: $\mathbb{E}[X_{2i}, X_{3i} | X_{1i}] = 0 \rightarrow \mathbb{E}[\varepsilon_i | X_{1i}] = 0$.
 - World 2: $\mathbb{E}[X_{2i}, X_{3i} | X_{1i}] \neq 0 \rightarrow \mathbb{E}[\varepsilon_i | X_{1i}] \neq 0$.
- If we are in World 1, you can use your random sample of (Y_i, X_{1i}) to learn something about β_1 . If we are in World 2, you cannot learn anything about β_1 , **unless you have access to other variables besides (Y_i, X_{1i})** .

World 1

- Imagine that you are in World 1:

$$\mathbb{E}[\text{avg. income pc} | \text{avg. class size}]$$

and

$$\mathbb{E}[\text{rainy days} | \text{avg. class size}]$$

do not depend on avg. class size: $\mathbb{E}[\varepsilon_i | X_{1i}] = 0$.

- Then, you can compute the OLS estimator for β_0 and β_1 using your random sample of (Y_i, X_{1i}) and it will be true that

$$\hat{\beta}_1^{OLS} \approx \mathbb{N}(\beta_1, \mathbb{V}[\hat{\beta}_1]).$$

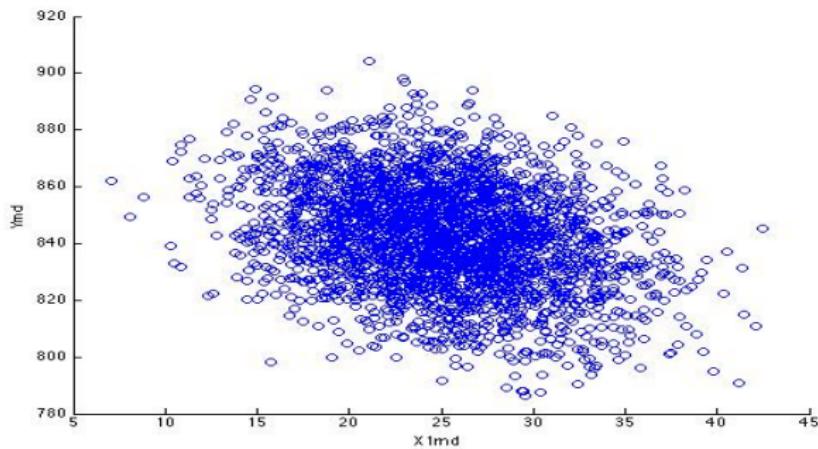
- If districts pick their class sizes randomly, $\mathbb{E}[\text{avg. income pc} | \text{avg. class size}]$ and $\mathbb{E}[\text{rainy days} | \text{avg. class size}]$ do not depend on avg. class size.
- **Random assignment is a particular example of World 1.**

World 1

- Imagine that the population from which we draw our random sample is such that:

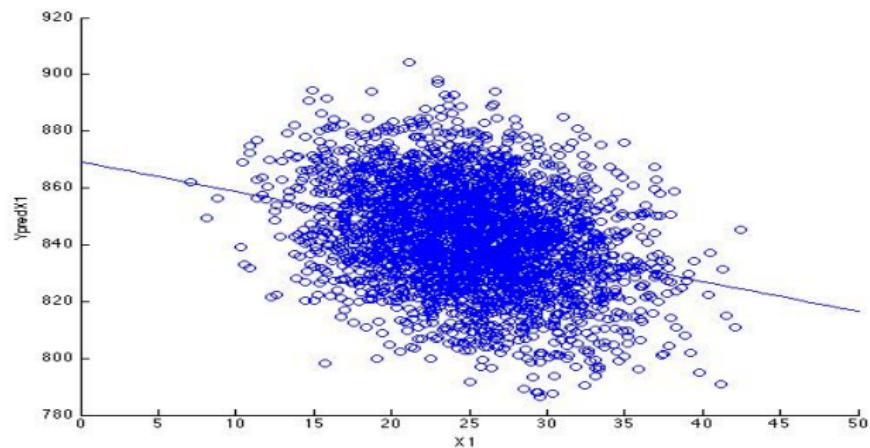
$$Y_i = 500 - 1X_{1i} + 7X_{2i} + 0.1X_{3i}$$

AND X_{1i} is independent of X_{2i} and X_{3i} . The population is:

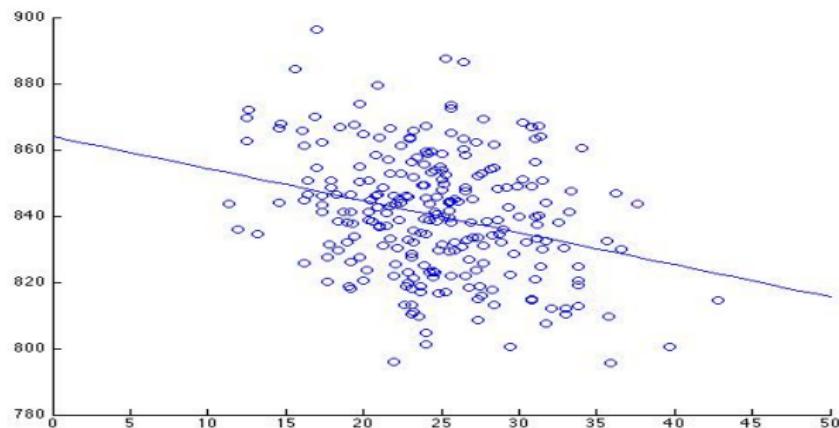


World 1

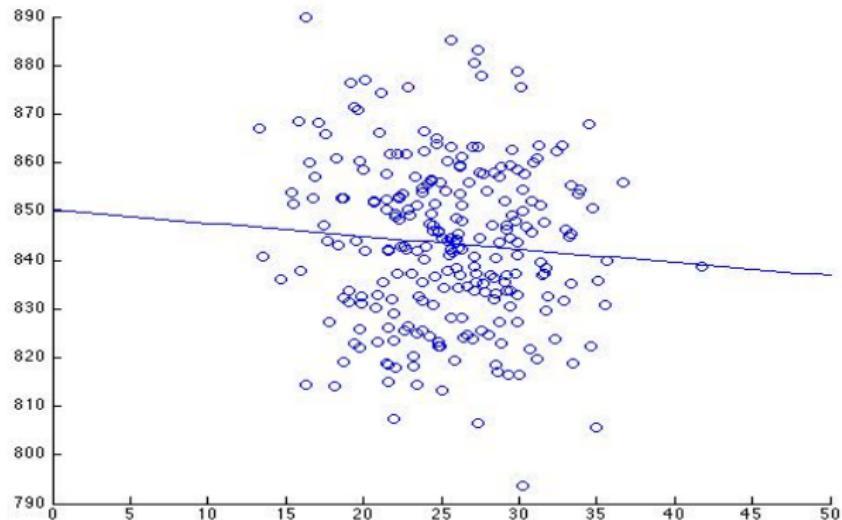
- If we compute the line that minimizes the squared differences in our population of (Y_i, X_{1i}) , the slope of this line is exactly $\beta_1 = -1$. This is the causal effect of interest.



- We will generally not see the whole population. We will usually see a random sample of the population. Our OLS estimate of the regression line is the line that minimizes the squared differences in our sample of (Y_i, X_{1i}) . The slope of this line is exactly our OLS estimate of β_1 . For one random sample of 300 districts, we obtain:

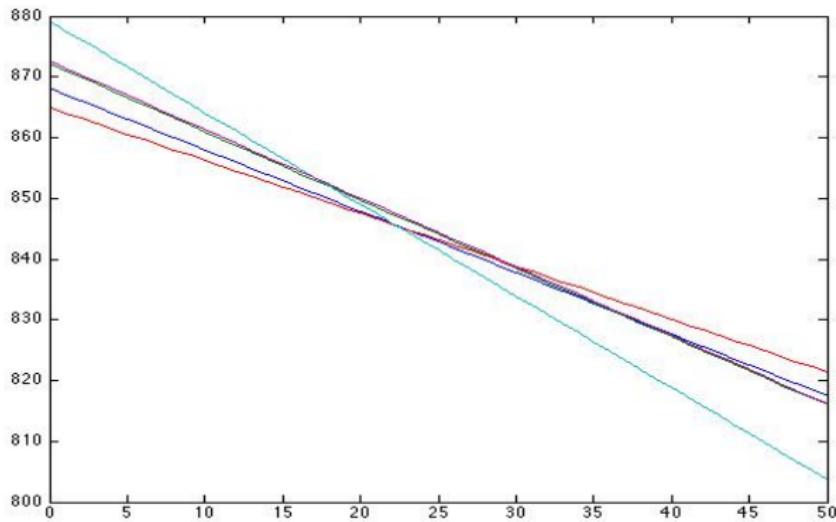


- For a second random sample of 300 districts, we obtain:



World 1

- If we take 5 different random samples of 300 districts and compute the OLS estimator of our regression line for each of them, we obtain:



- Our 5 estimates of β_1 are: $-1.01, -1.12, -0.87, -1.55, -1.13$ (i.e. not exactly equal to -1 but very close).

- If, instead of only 5 samples, I had taken many many samples of size 300 from the population of 13,500 school districts and had computed the distribution of my estimates $\hat{\beta}_1$, I would have obtained a distribution that is very close to normal.
- In general, you will only have access to one random sample from the population of interest. Therefore, your OLS estimator will give you just one of the regression lines in the previous figure.
- The key finding is that: as long as the population from which we take our random sample is such that the average class size, X_{1i} , is mean independent of any other factor or variable affecting grades, Y_i , then your OLS estimator will give you valuable information about the causal effect of interest: β_1 .

- Imagine now that, in the US, richer school districts (high values of X_{2i}) tend to have smaller class sizes (low values of X_{1i}).
- In this case, it happens that both X_{2i} and X_{1i} are negatively correlated and, therefore,

$$\mathbb{E}[\text{avg. income pc} | \text{avg. class size}]$$

will actually depend on the average class size.

- If we have a random sample of (Y_i, X_{1i}) from a population in which X_{1i} is correlated with some other variable X_{2i} that also has a causal impact on Y_i , what can we learn about the causal effect of X_{1i} on Y_i (i.e. β_1) from our random sample of (Y_i, X_{1i}) ? **Basically nothing.**
- In the next lecture, we will see how, in the case in which X_{1i} and X_{2i} are correlated, we can use data on X_{2i} (jointly with data on (Y_i, X_{1i})) to recover a consistent estimate of β_1 .

World 2

- Imagine that the population from which we draw our random sample is such that:

$$Y_i = 500 - 1X_{1i} + 7X_{2i} + 0.1X_{3i}$$

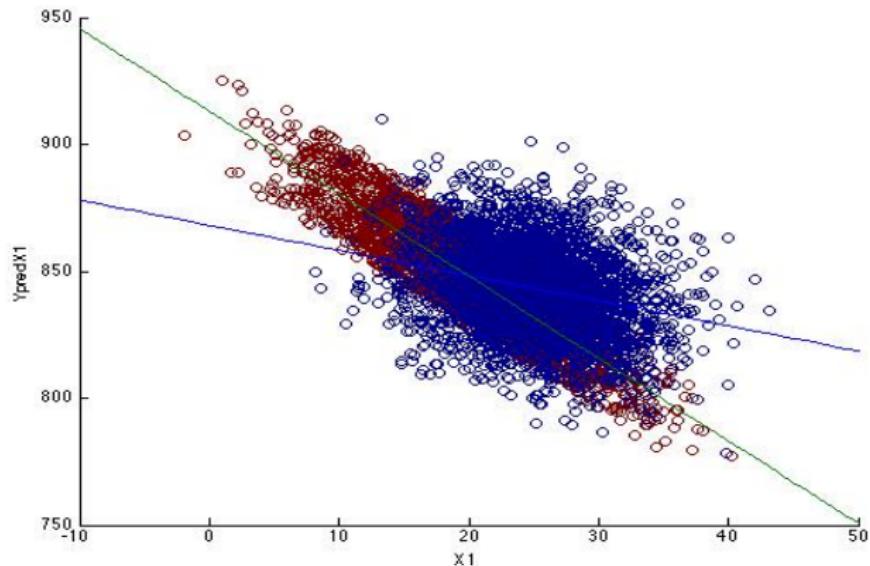
AND

$$X_{1i} \sim \mathbb{N}(170 - 3X_{2i}, 1)$$

and both X_{1i} and X_{2i} are independent of X_{3i} .

- What can we learn about β_1 ?

World 2



Blue dots and blue line: population and conditional expectation in World 1.

Red dots and green line: population and conditional expectation in World 2.

The slope of the blue line is β_1 . The slope of the green line is NOT equal to β_1 .

- If we only have access to a random subset of the red dots, what can we learn about the slope of the blue line?
- Not much. The OLS estimator in our sample will verify

$$\hat{\beta}_1^{OLS} \approx \mathbb{N}(\gamma, \mathbb{V}[\hat{\beta}_1^{OLS}])$$

where γ denotes the slope of the green line in the previous figure. But, as we saw in the previous figure, $\gamma \neq \beta_1$.

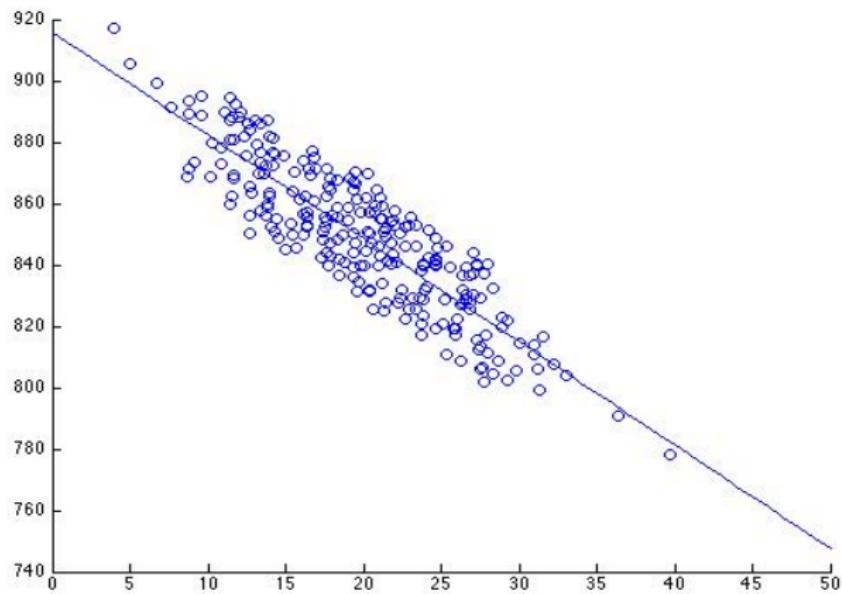
- The OLS estimator, $\hat{\beta}_1$, in a random sample of (Y_i, X_{1i}) in a population in which X_{i1} is correlated with other factor affecting Y_i will still be a great estimator of the slope of conditional expectation

$$\mathbb{E}[Y_i | X_{1i}]$$

in the population from which the sample is drawn. The problem is that this slope is not equal to the treatment or causal effect of class size on grades (which is the parameter that the Secretary of Education cares about!).

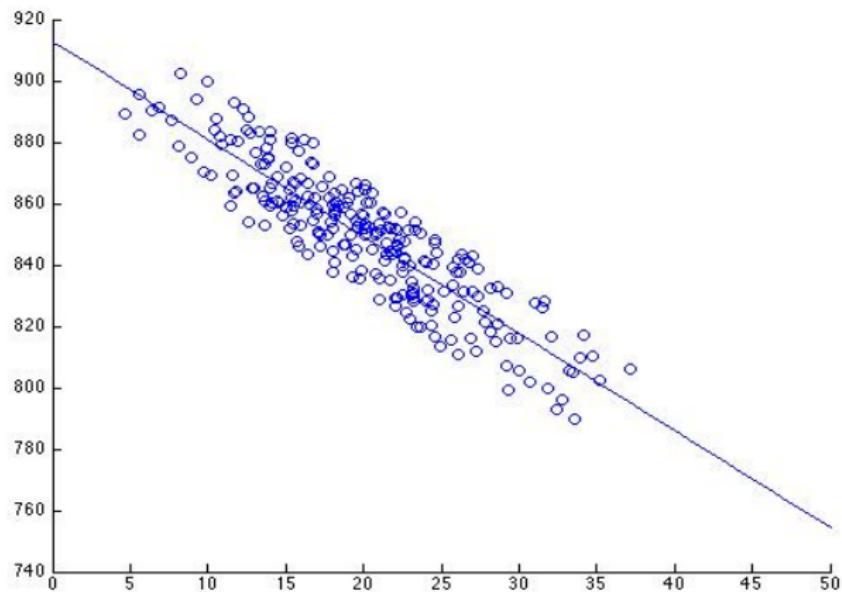
World 2

- This is a random sample of 300 districts from World 2.



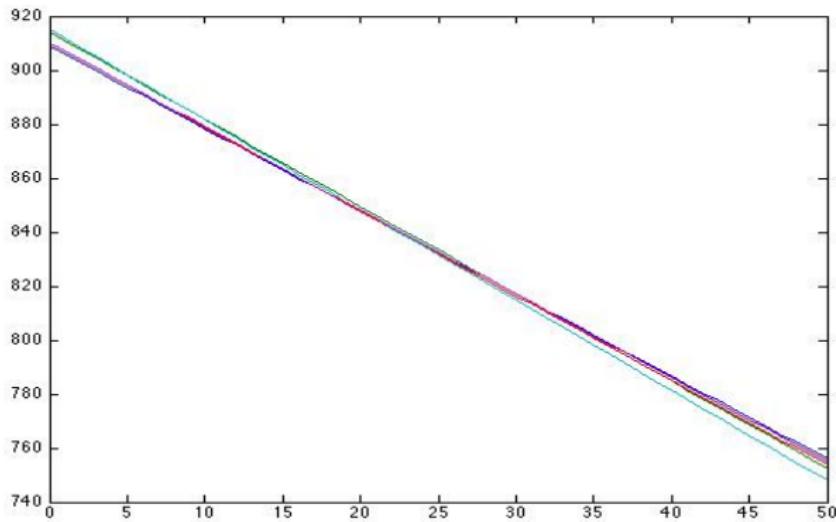
World 2

- This is a second random sample of 300 districts from World 2.



World 2

- If we take 5 different random samples of 300 districts and compute the OLS estimator of our regression line for each of them, we obtain:



- The slope of these lines is: $-3.05, -3.23, -3.12, -3.34, -3.10$ (i.e. quite close to -3 and far away from -1).

Summary

- Independently of whether we are in World 1 or in World 2, as long as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i,$$

with β_1 denoting the causal effect of X_i on Y_i , a random sample of (Y_i, X_i) and the OLS estimator computed on this sample will **ALWAYS** give you useful information about the slope of conditional expectation

$$\mathbb{E}[Y_i | X_{1i}]$$

- However, **ONLY IF** we are in World 1 (**ONLY if** $\mathbb{E}[\varepsilon_i | X_{1i}] = 0$), it will be true that a random sample of (Y_i, X_i) and the OLS estimator computed on this sample will give us useful information about the causal effect of X_{1i} on Y_i .

Additional Note: Heteroskedasticity

- What happens if, in the population of interest, $\mathbb{E}[\varepsilon_i | X_{1i}]$ does not depend on X_{1i} but $\mathbb{V}[\varepsilon_i | X_{1i}]$ depends on X_{1i} ?
- Note 1: if $\mathbb{V}[\varepsilon_i | X_{1i}]$, then X_{1i} has not been randomly assigned in the population of interest.
- Note 2: even if X_{1i} has not been randomly assigned in the population of interest, the fact that it is still true that $\mathbb{E}[\varepsilon_i | X_{1i}]$ does not depend on X_{1i} implies that the OLS estimator $\hat{\beta}_1$ will still satisfy

$$\hat{\beta}_1 \approx \mathbb{N}(\beta_1, \mathbb{V}[\hat{\beta}_1]),$$

where β_1 is the causal effect of interest. However, we need to compute $\mathbb{V}[\hat{\beta}_1]$ in a way that it takes into account that $\mathbb{V}[\varepsilon_i | X_{1i}]$ depends on X_{1i} : we need to estimate the variance of $\hat{\beta}_1$ using the heteroskedasticity-robust estimator.

WWS 507c: Quantitative Analysis

Lecture 13: Causal Effect and Treatment Variables:
Linear Regression with Multiple Regressors.

Princeton University

November 16, 2014

MOTIVATING EXAMPLE

Example

- The causal relationship between grades, Y_i , average class size (var. 1), average income per capita (var. 2), and number of rainy days (var. 3), is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i},$$

with $\beta_0 = 500$, $\beta_1 = -1$, $\beta_2 = 7$, $\beta_3 = 0.1$, and each variable X measures differences with respect to its population mean.

- School districts with higher income per capital tend to have lower average class sizes. In particular,

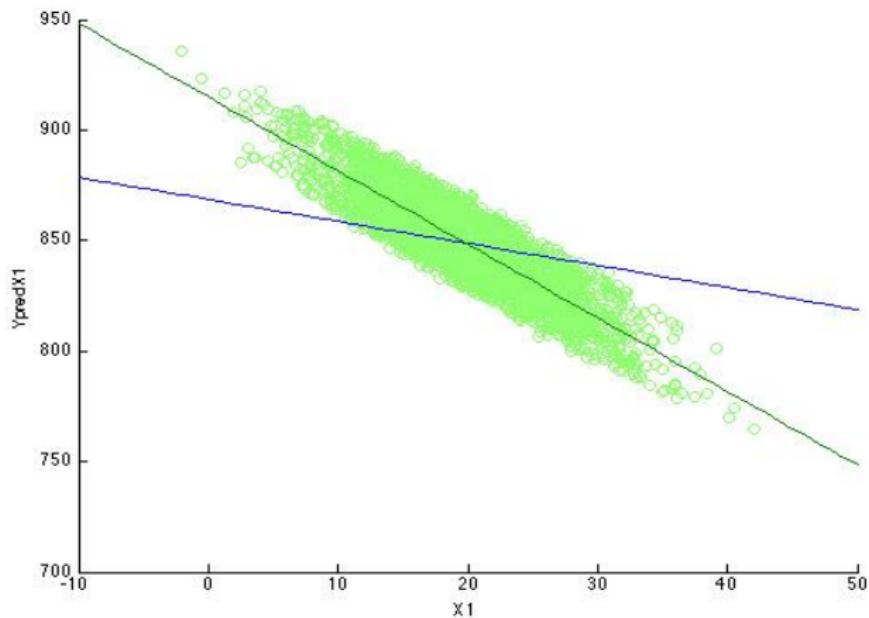
$$X_1 \sim \mathbb{N}(170 - 3X_2, 1),$$

This implies that $\text{corr}(X_{1i}, X_{2i}) = -0.98 \leftrightarrow \mathbb{E}[X_2 | X_1] \neq 0$.

- Number of rainy days across districts is not correlated with either average class size or average income per capita:

$$\mathbb{E}[X_3 | X_1, X_2] = 0, \quad \leftrightarrow \quad \text{corr}(X_3, X_1) = 0, \text{ and } \text{corr}(X_3, X_2) = 0.$$

Example



Example

- Explanation of previous figure:
- Green dots: distribution of (X_{1i}, Y_i) in the population.
- Green line: $\mathbb{E}[Y|X_1]$ in the population. In our particular example, the slope of $\mathbb{E}[Y|X_1]$ is -3.25 . The slope of the green line is -3.25 .
- Blue line: it captures the causal relationship between X_{1i} and Y_i :
$$Y_i = \beta_0 + \beta_1 X_{1i}$$
. Therefore, the slope of this line is the causal effect of interest, β_1 . In our particular example, β_1 is equal to -1 .

Remember that the blue line would also capture the conditional expectation $\mathbb{E}[Y|X_1]$ in a world in which $\mathbb{E}[\varepsilon|X_1] = 0$, with

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 X_{1i} = \beta_2 X_{2i} + \beta_3 X_{3i}.$$

If $\mathbb{E}[\varepsilon|X_1] = 0$, then the slope of the conditional expectation $\mathbb{E}[Y|X_1]$ and the slope of the line capturing the causal effect of X_1 on Y , β_1 , are the same.

QUESTIONS FOR TODAY

What are we Doing Today?

- Given the setting described in slide 2, we will answer two questions:
- Question 1:** How does β_1 compare to the slope of $\mathbb{E}[Y|X_1]$? Is it larger or smaller? Why?
- In words, question 1 asks: in a world in which the treatment variable of interest X_1 is correlated with an omitted variable X_2 that affects the outcome variable of interest, Y , how does the slope of the expectation of Y conditional on X_1 compares to the causal effect of X_1 on Y ?
- Question 2:** If our random sample, besides data on test scores (outcome) and average class size (treatment), were to include information on average income per capita in each district (variable correlated with treatment), would it be possible to use this random sample

$$\{(Y_i, X_{1i}, X_{2i}), i = 1, \dots, N\}$$

to find an estimator $\hat{\beta}_1$ that is consistent for the causal effect of interest, β_1 ?

Question 1: Causal Effect vs. Slope of $\mathbb{E}[Y|X_1]$

- Let's think about the main features of particular example:
 - causal effect of avg. income p.c., X_2 , on student grades, Y , is positive: $\beta_2 = 7$.
 - average income per capita, X_2 , and average class size, X_1 , are negatively correlated: $\text{corr}(X_1, X_2) = -0.98$.
 - average class size, X_1 , is not correlated with any other factor/variable affecting output, Y : $\text{corr}(X_1, X_3) = 0$.
- Under these conditions, we obtain that the slope of $\mathbb{E}[Y|X_1]$ is smaller than the causal effect of X_1 on Y :

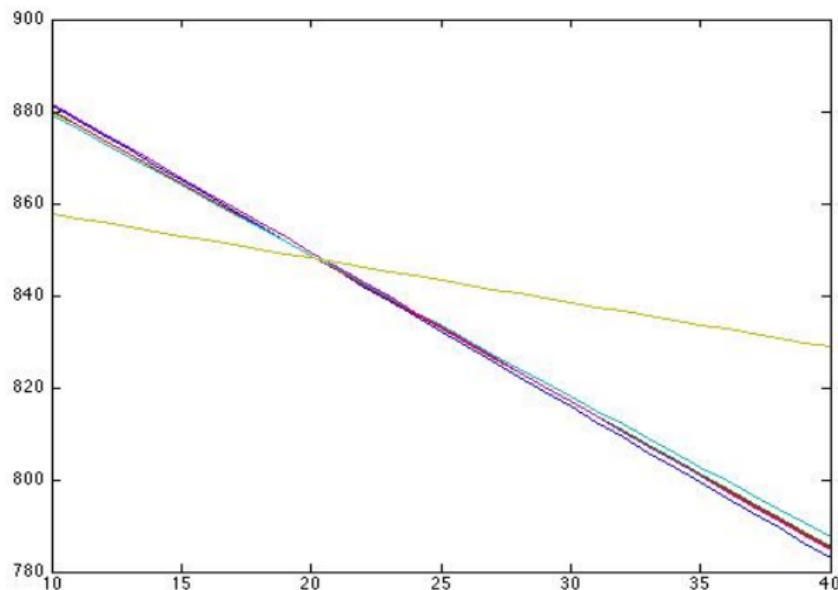
$$-3.25 < -1.$$

- Given that, under random sampling, the expectation of the OLS estimator for the effect of X_1 on Y is the same as the slope of the expectation of Y conditional on X_1 ,

$$\mathbb{E}[\hat{\beta}_1] = -3.25 \quad (= \text{slope of } \mathbb{E}[Y|X_1])$$

this implies that $\mathbb{E}[\hat{\beta}_1] < \beta_1$ ($=$ causal effect of X_1 on Y).

Question 1: Causal Effect vs. Slope of $\mathbb{E}[Y|X_1]$



β_1 : slope of the yellow line.

$\mathbb{E}[\hat{\beta}_1]$: average of the slopes of the blue, red, green... lines.

Question 1: Causal Effect vs. Slope of $\mathbb{E}[Y|X_1]$

- Remember that the difference between the expectation of an estimator and the parameter that we are trying to estimate is called **bias**.

$$\text{bias of } \hat{\beta}_1 = \mathbb{E}[\hat{\beta}_1] - \beta_1.$$

- The bias in the OLS estimator of the causal effect of some variable X_1 on an outcome variable Y that is due to the correlation of X_1 with an unobserved variable X_2 affecting Y is known as **omitted variable bias**.
- The bias of $\hat{\beta}_1$ as an estimator of the causal effect of X_1 on Y is equal to 0 only if:
 - there is no other variable besides X_1 affecting Y (in our ex. $\beta_2 = \beta_3 = 0$); **OR**,
 - the expectation of any other variable affecting the outcome Y conditional on the treatment variable of interest, X_1 , does not depend on X_1 . In our example:

$$\mathbb{E}[X_2|X_1] = \mathbb{E}[X_3|X_1] = 0.$$

Question 1: Causal Effect vs. Slope of $\mathbb{E}[Y|X_1]$

- **General expression for omitted variable bias:**

$$\hat{\beta}_1 \xrightarrow{N \rightarrow \infty} \beta_1 + \text{corr}(X_1, \varepsilon) \frac{\sigma_\varepsilon}{\sigma_{X_1}} = \beta_1 + \frac{\text{cov}(\varepsilon, X_1)}{\sigma_{X_1}^2}$$

- In our particular example:

$$\varepsilon = \beta_2 X_2 + \beta_3 X_3,$$

therefore, with the assumption that $\text{cov}(X_3, X_1) = 0$, we obtain

$$\begin{aligned}\text{cov}(\varepsilon, X_1) &= \text{cov}(\beta_2 X_2 + \beta_3 X_3, X_1) = \beta_2 \text{cov}(X_2, X_1) + \beta_3 \text{cov}(X_3, X_1) \\ &= \beta_2 \text{cov}(X_2, X_1)\end{aligned}$$

- Given that $\beta_2 > 0$ and $\text{cov}(X_2, X_1) < 0$, our general expression for the probability limit of $\hat{\beta}_1$ indicates that: $\text{plim}(\hat{\beta}_1) < \beta_1$ (this is precisely what we found in our simulations in slide 7).

Question 2: What can we do if our dataset includes X_2 ?

- Imagine now that you observe a random sample of school districts that, for each school district, includes information on
 - test scores: outcome variable;
 - average class size: treatment variable;
 - average income per capita: additional determinant of outcome variable.
- Now the only unobserved factor affecting test scores is number of rainy days.
- The number of rainy days is mean independent of the two observed covariates: independent of average class size and average income per capita.
- Therefore:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

with $\varepsilon_i = \beta_3 X_{3i}$ and

$$\mathbb{E}[\varepsilon_i | X_{1i}, X_{2i}] = 0 \quad \rightarrow \quad \mathbb{E}[\beta_3 X_{3i} | X_{1i}, X_{2i}] = 0.$$

Question 2: What can we do if our dataset includes X_2 ?

- Before describing the properties of the OLS estimator in this setting, let's update our definition of the OLS estimator to the case in which we observe more than one covariate.
- Now we have three parameters to estimate: $(\beta_0, \beta_1, \beta_2)$. Accordingly, we will have an OLS estimator for each of them: $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$.
- As in the case in which we only had one regressor, the OLS estimator is again the set of values that minimizes the sum of squared differences between each value of Y_i in the sample and its predicted value according to the model:

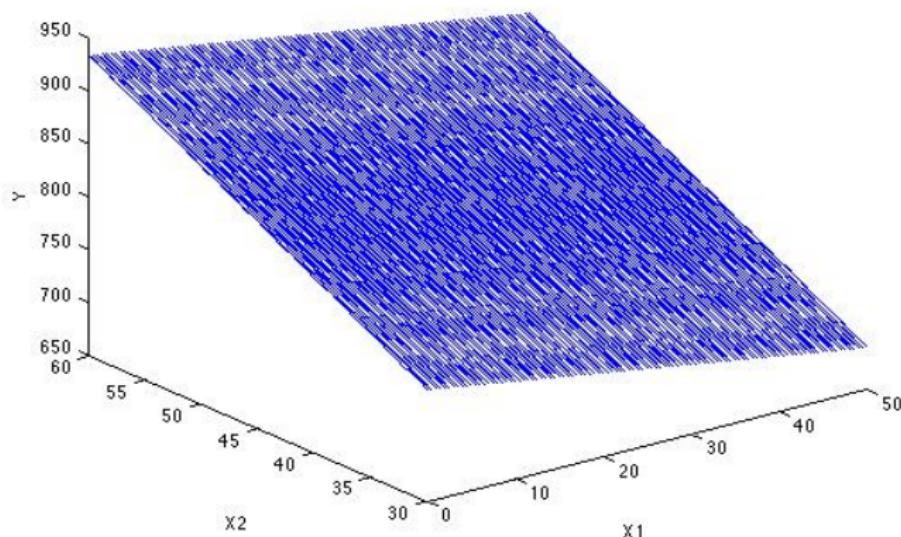
$$\text{predicted value} = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

- In mathematical terms, the OLS estimator $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is defined as:

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \underset{b_0, b_1, b_2}{\operatorname{argmin}} \sum_{i=1}^N (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

Question 2: What can we do if our dataset includes X_2 ?

- In our particular, example, the regression of Y on X_1 and X_2 is:



- The regression line becomes a regression plane!

Question 2: What can we do if our dataset includes X_2 ?

- Properties of the OLS estimator.
- **IF**
 - $\mathbb{E}[\varepsilon|X_1, X_2] = 0$
 - $\{(Y_i, X_{1i}, X_{2i}); i = 1, \dots, N\}$ is a random sample from the population of interest.
 - Y , X_1 , and X_2 have finite fourth moments.
 - $\text{corr}(X_1, X_2) \neq 1$ and $\text{corr}(X_1, X_2) \neq -1$ (i.e. **no perfect multicollinearity**)
- **THEN**
 - $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ is unbiased, consistent and asymptotically normal for $(\beta_0, \beta_1, \beta_2)$.
- In particular,

$$\frac{\hat{\beta}_1 - \beta_1}{\mathbb{V}[\hat{\beta}_1]} \approx \mathbb{N}(0, 1).$$

Question 2: What can we do if our dataset includes X_2 ?

- Summary:
- We are in a world in which Y_1 is *caused* by X_1 , X_2 , and X_3 and
 - X_1 and X_2 are correlated;
 - X_3 is uncorrelated with both X_1 and X_2
- If we only have a random sample of Y and X_1 (i.e. the error term ε includes both X_2 and X_3) then

$$\mathbb{E}[\varepsilon|X_1] \neq 0,$$

and, therefore, $\hat{\beta}_1$ is NOT an unbiased nor consistent estimator of β_1 .

- If we have a random sample of Y , X_1 , and X_2 (i.e. the error term ε includes only X_3) then

$$\mathbb{E}[\varepsilon|X_1, X_2] = 0 \quad \xrightarrow{\text{by the LIE}} \quad \mathbb{E}[\varepsilon|X_1] = 0,$$

and, therefore, $\hat{\beta}_1$ is an unbiased and consistent estimator of the causal effect of interest β_1 .

Question 2: What can we do if our dataset includes X_2 ?

- We can also see this graphically:

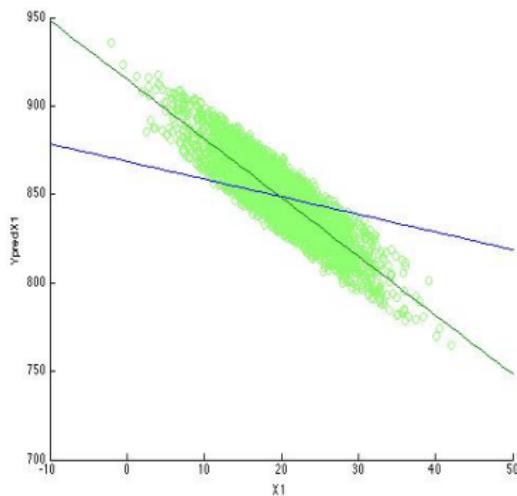


Figure : Only observe X_1

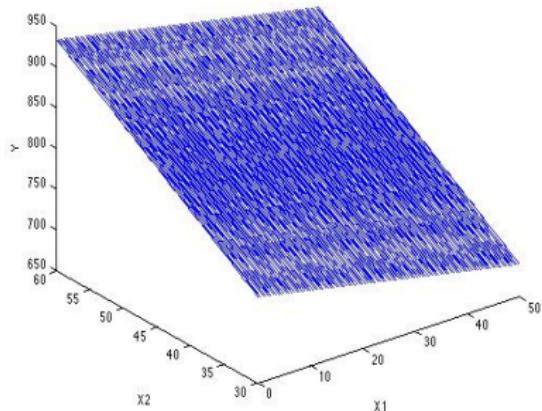


Figure : Observe X_1 and X_2

The slope of the plane along the X_1 dimension (right figure) is the same as the slope of the blue line in the left figure!

Question 2: What can we do if our dataset includes X_2 ?

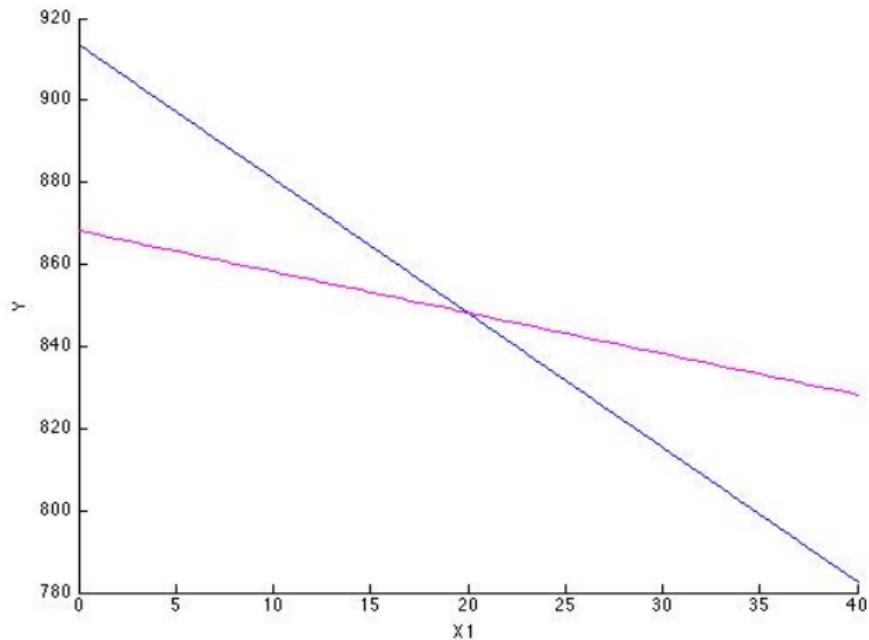
- How is it possible that including X_2 in the regression allows us to estimate the causal effect of X_1 on Y ?
- Let's think about the variation in the data that helps us measure/estimate the causal effect of X_1 on Y .
- When we only observe Y and X_1 , we quantify our estimate of the causal effect of X_1 on Y by comparing observations that take different values of X_1 (e.g. observations with $X_1 = 20$ and $X_1 = 25$) and measuring how different Y is for those observations **on average**.
- The problem is that those two groups of observations will have different values of Y on average not only because their values of X_1 are different, but also because, **on average**, those observations with lower values of X_1 tend to have higher values of X_2 , and X_2 also has an impact on Y .
- Therefore, when we regress Y on X_1 , our comparison of average values of Y for different values of X_1 captures both the causal effect that X_1 has on Y as well as the effect that X_2 has on Y .

Question 2: What can we do if our dataset includes X_2 ?

- In order to estimate the causal effect of X_1 on Y , we would like that:
 - ➊ in the population interest, observations with different values of X_1 do not differ **on average** on their values of X_2 . Mathematically, we express this as " $E[X_2|X_1]$ does not depend on X_1 ". This is precisely what is attained when we randomly assign different values of X_2 to the different individuals in the population from which we draw our sample.
 - ➋ or, if we cannot avoid X_2 being correlated with X_1 in the population of interest (i.e. we cannot randomize our treatment variable), we would like to quantify or measure the effect of X_1 on Y by comparing the values of Y for observations that differ in their values of X_1 **but have identical values of X_2** .
- In statistics, we denote the procedure in number 2 as to **control for X_2 when estimating the effect of X_1 on Y** or to **partial out the effect of X_2 when estimating the effect of X_1 on Y** .
- This is precisely what the OLS estimator of β_1 does when we include both X_1 and X_2 in our regression!

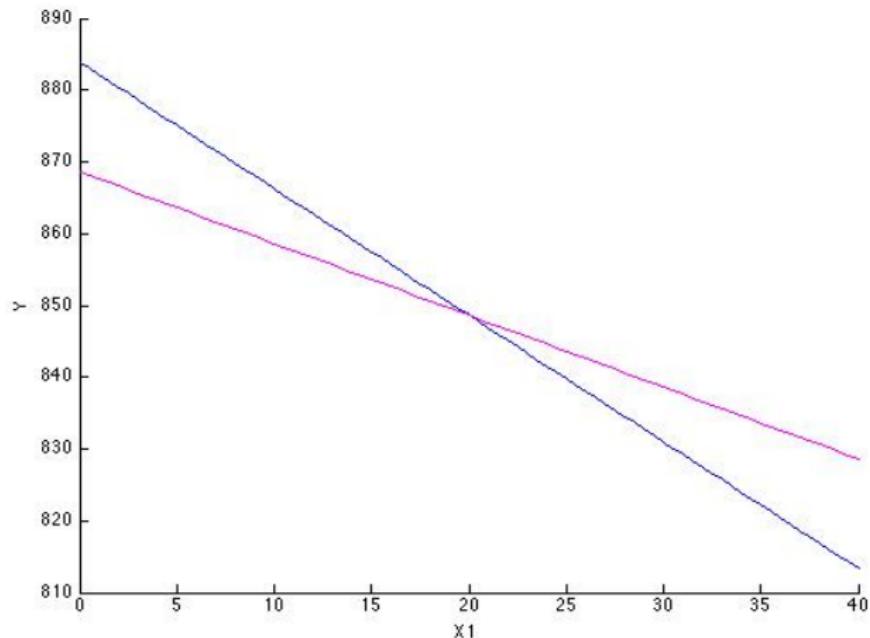
What does “to control for” mean?

Regression line and causal effect line when regression line is estimated on the whole population (i.e. without any restriction on X_2)



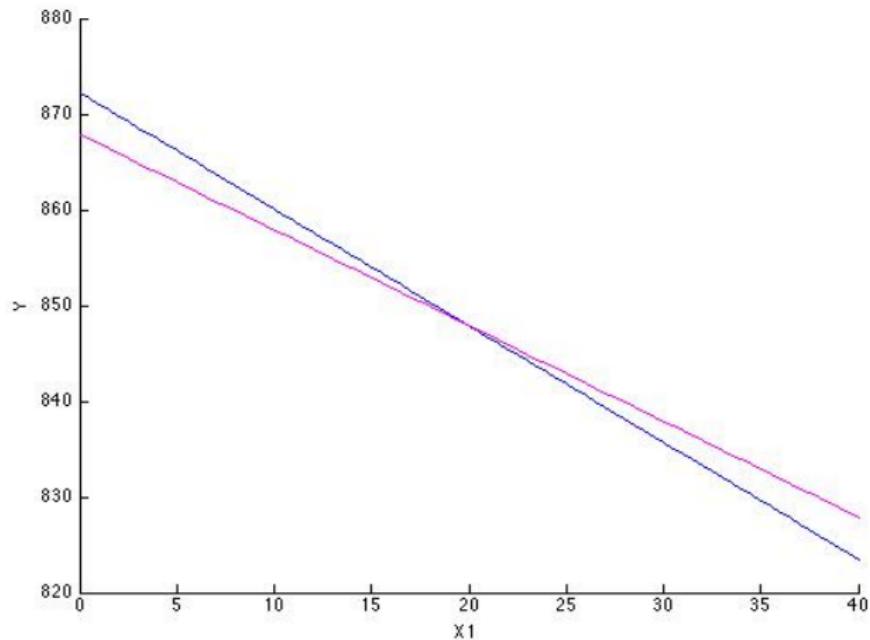
What does “to control for” mean?

Regression line and causal effect line when regression line is estimated on observations i such that (percentile 40 of X_2) $< X_{2i} <$ (percentile 60 of X_2).



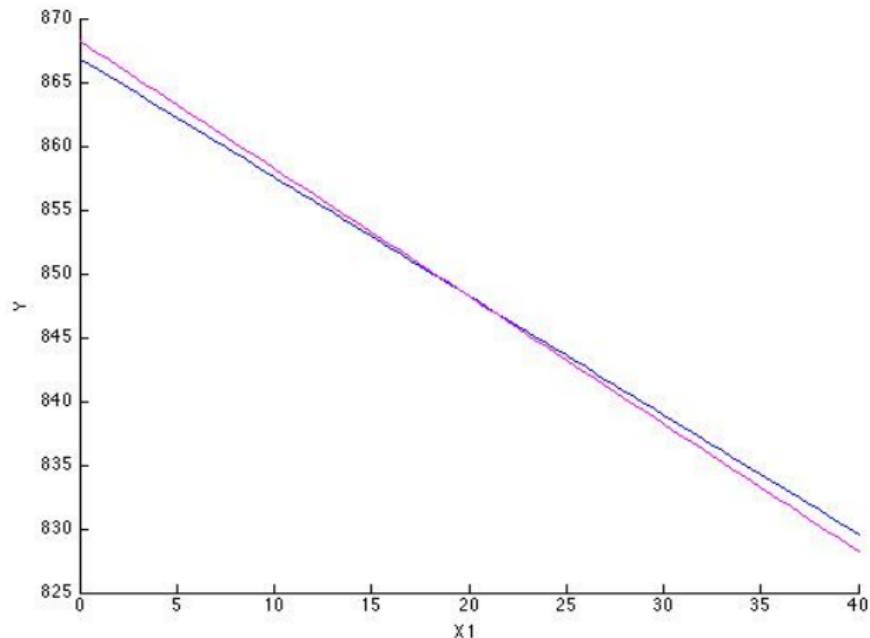
What does “to control for” mean?

Regression line and causal effect line when regression line is estimated on observations i such that (percentile 45 of X_2) $< X_{2i} <$ (percentile 55 of X_2).



What does “to control for” mean?

Regression line and causal effect line when regression line is estimated on observations i such that (percentile 49 of X_2) $< X_{2i} <$ (percentile 51 of X_2).



FRISCH-WAUGH-LOVELL THEOREM

What does “to partial out” mean?

- An alternative way of computing the OLS estimator of β_1 in a regression with multiple covariates.
- Again, to fix ideas, let's imagine that we want to compute the OLS estimator of β_1 in the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

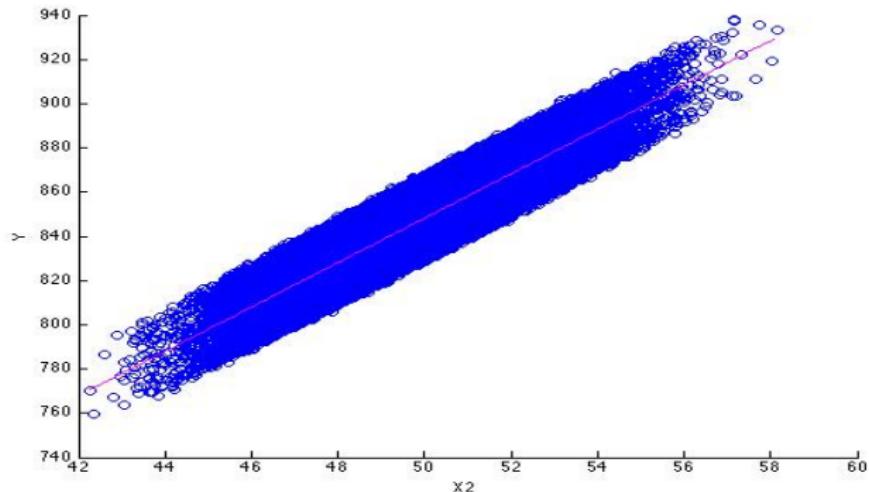
- As we saw in slide 14, one way to compute the OLS estimator of β_1 is to solve

$$(\hat{\beta}_1, \hat{\beta}_1, \hat{\beta}_2) = \underset{b_0, b_1, b_2}{\operatorname{argmin}} \sum_{i=1}^N (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

- An alternative way to recover $\hat{\beta}_1$ is the following:
 - ① Regress Y on a constant and X_2 . Define the OLS residuals as $\hat{\varepsilon}_i^Y$.
 - ② Regress X_1 on a constant and X_2 . Define the OLS residuals as $\hat{\varepsilon}_i^X$.
 - ③ Regress $\hat{\varepsilon}_i^Y$ on a constant and $\hat{\varepsilon}_i^X$. The OLS estimator of the coefficient on $\hat{\varepsilon}_i^X$ is exactly $\hat{\beta}_1$.

What does “to partial out” mean?

Step 1:

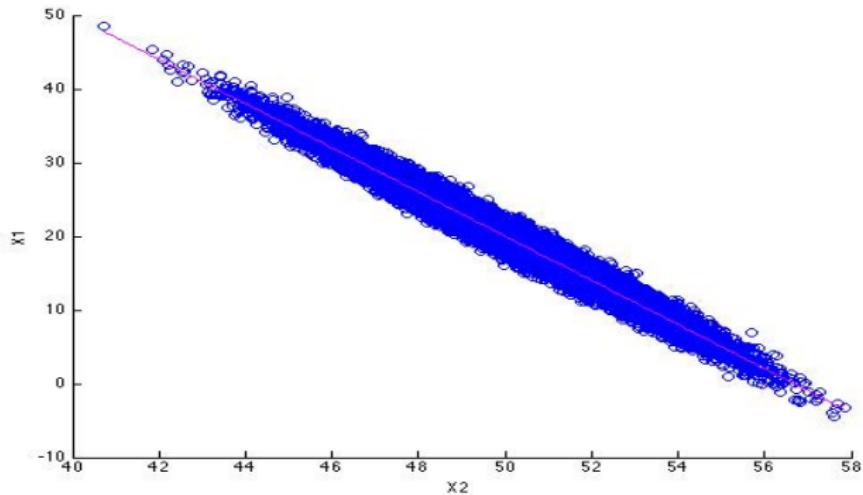


$$Y = \hat{\gamma}_0 + \hat{\gamma}_1 X_{2i} + \hat{\varepsilon}_i^Y, \quad \text{with } \hat{\gamma}_0 = 346 \text{ and } \hat{\gamma}_1 = 10.043$$

$(\hat{\gamma}_0, \hat{\gamma}_1)$ are just OLS estimators. They have no causal interpretation.

What does “to partial out” mean?

Step 2:

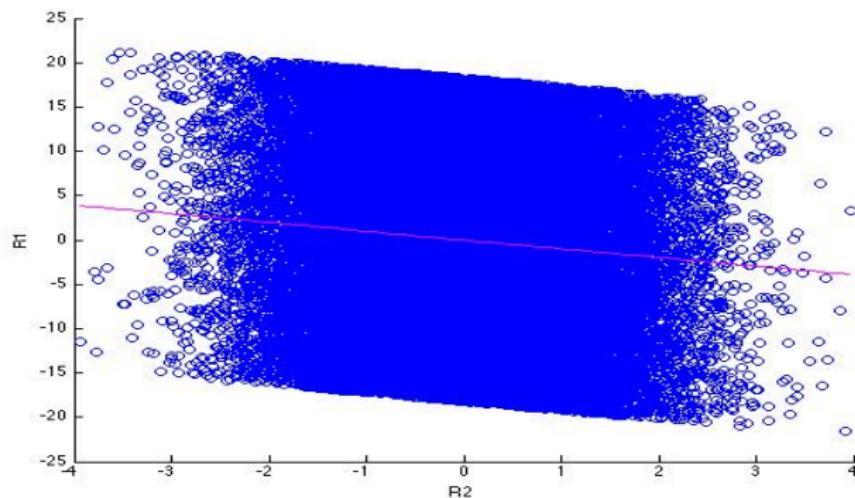


$$X_{1i} = \hat{\delta}_0 + \hat{\delta}_1 X_{2i} + \hat{\varepsilon}_i^X, \quad \text{with } \hat{\delta}_0 = 170 \text{ and } \hat{\delta}_1 = -2.99$$

$(\hat{\delta}_0, \hat{\delta}_1)$ are just OLS estimators. They have no causal interpretation.

What does “to partial out” mean?

Step 3:



$$\hat{\varepsilon}_i^Y = \hat{\beta}_0 + \hat{\beta}_1 \hat{\varepsilon}_i^X + \varepsilon_i, \quad \text{with } \hat{\beta}_0 = 0 \text{ and } \hat{\beta}_1 = -0.99 \approx -1$$

and $\hat{\beta}_1$ is a consistent estimator of the causal effect of X_1 on Y !!!!!

OBTAINING CONSISTENT ESTIMATES OF SOME (BUT NOT ALL) PARAMETERS

Consistent Estimates of Some Parameters

- As we saw in slide 16, in a regression model in which we regress Y on a constant and two covariates, (X_1, X_2) ,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

a **necessary assumption** so that the OLS estimator $(\hat{\beta}_1, \hat{\beta}_2)$ is unbiased and consistent for (β_1, β_2) is that

$$\mathbb{E}[\varepsilon | X_1, X_2] = \mathbb{E}[Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 | X_1, X_2] = 0.$$

- In words: if we want to use a random sample of districts with data on test scores, class size, and average income to recover the causal effect of **both** class size and average income on test scores, then we need to assume: (a) the causal relationship of **both** class size, X_{1i} , and average income, X_{2i} , on test scores, Y_i , is linear; (b) every other variable that might affect test scores is mean independent of **both** class size and income.

Consistent Estimates of Some Parameters

- Given a regression of Y on X_1 and X_2 , which assumptions do we need to impose if we only care about obtaining a consistent estimator for β_1 ?
 - Which assumptions do we need to impose if we are only interested in obtaining a consistent estimator of the causal effect of class size on grades?
- In this case, we need to impose the following assumptions:
 - $\mathbb{E}[\varepsilon|X_1, X_2] = \mathbb{E}[\varepsilon|X_2]$
 - $\{(Y_i, X_{1i}, X_{2i}); i = 1, \dots, N\}$ is a random sample.
 - Y , X_1 , and X_2 have finite fourth moments.
 - $\text{corr}(X_1, X_2) \neq 1$ and $\text{corr}(X_1, X_2) \neq -1$ (i.e. **no perfect multicollinearity**)
- Assumptions 2, 3, and 4 are identical to those needed for both $\hat{\beta}_1$ and $\hat{\beta}_2$ to be unbiased and consistent for β_1 and β_2 .
- Assuming that $\mathbb{E}[\varepsilon|X_1, X_2] = \mathbb{E}[\varepsilon|X_2]$ is **weaker** than assuming that $\mathbb{E}[\varepsilon|X_1, X_2] = 0$. If $\mathbb{E}[\varepsilon|X_1, X_2] = \mathbb{E}[\varepsilon|X_2]$, but $\mathbb{E}[\varepsilon|X_1, X_2] \neq 0$, then $\hat{\beta}_1$ is a consistent estimator of β_1 but $\hat{\beta}_2$ is not a consistent estimator of β_2 .

Consistent Estimates of Some Parameters

- How is it possible that $\mathbb{E}[\varepsilon|X_1, X_2] = \mathbb{E}[\varepsilon|X_2]$ but $\mathbb{E}[\varepsilon|X_1, X_2] \neq 0$?
- Let's come back to our example to get some intuition. Remember that:

$$Y_i = 500 - 1X_{1i} + 7X_{2i} + 0.1X_{3i}$$

with X_1 = class size, X_2 = income per capita, and X_3 = number rainy days.

- Imagine that income per capita is higher in districts with more rainy days per year (e.g. I need to pay you more for you to agree to move to a place with bad weather) and that higher income per capita districts have smaller class sizes (e.g. higher income pc implies higher tax revenues and school budgets). In this case, $\text{cov}(X_{1i}, X_{3i}) < 0$ and $\text{cov}(X_{2i}, X_{3i}) > 0$ and $\text{cov}(X_{2i}, X_{1i}) < 0$.
- However, if the number of rainy days only affects class sizes through income per capita (i.e. among the subpopulation of districts with identical income per capita, there is no correlation between number of rainy days and class sizes) then it will be true that

$$\mathbb{E}[X_3|X_1, X_2] = \mathbb{E}[X_3|X_2].$$

CONTROL VARIABLES

Control Variables

- Let's go back to our original example.
- The causal relationship between grades, Y_i , average class size (var. 1), average income per capita (var. 2), and number of rainy days (var. 3), is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i},$$

with $\beta_0 = 500$, $\beta_1 = -1$, $\beta_2 = 7$, $\beta_3 = 0.1$, and each variable X measures differences with respect to its population mean.

- School districts with higher income per capital tend to have lower average class sizes. In particular,

$$X_1 \sim N(170 - 3X_2, 1),$$

This implies that $\text{corr}(X_{1i}, X_{2i}) = -0.98 \leftrightarrow \mathbb{E}[X_2 | X_1] \neq 0$.

- Number of rainy days across districts is not correlated with either average class size or average income per capita:

$$\mathbb{E}[X_3 | X_1, X_2] = 0, \leftrightarrow \text{corr}(X_3, X_1) = 0, \text{ and } \text{corr}(X_3, X_2) = 0.$$

Control Variables

- Up until now we have seen that:
 - If we have a random sample of (Y_i, X_{1i}) and regress Y on X_1 , then $\hat{\beta}_1$ is not a consistent estimator of the causal effect of X_1 on Y .
 - If we have a random sample of (Y_i, X_{1i}, X_{2i}) and regress Y on X_1 and X_{2i} , then $\hat{\beta}_1$ is a consistent estimator of the causal effect of X_1 on Y , and $\hat{\beta}_2$ is a consistent estimator of the causal effect of X_2 on Y .
- Now we will consider a third case in which we do not observe X_{2i} and we can still obtain a consistent estimator of β_1 .
- Imagine that we do not observe average income per capita X_{2i} for each district but, instead, we observe either the percentage of students receiving a free or subsidized school lunch.
- This variable will be a particular example of what we call **control variable** and we will denote it as W_{2i} .

Control Variables

- What is a control variable?
- It is a variable that is included to hold constant factors that, if neglected, could lead the estimated causal effect of interest to suffer from omitted variable bias.
- In our particular example, a control variable is a variable such that, if we control for it, the mean dependency between class size and income per capita becomes 0.
- A control variable is a variable such that, if we limit ourselves to observations that have the same value of this control variable, then knowing the value of the treatment effect of interest (i.e. class size) does not give any additional information about the value of the omitted variable (i.e. income per capita)

$$\mathbb{E}[X_{2i} | W_i, X_{1i}] = \mathbb{E}[X_{2i} | W_i].$$

Control Variables

- Why are control variables useful?
- They allow us to use OLS estimators to recover consistent estimates for the causal effect of some variable X_1 on some variable Y even when there is some unobservable variable X_2 that is both correlated with X_1 and has a causal effect on Y .
- In particular, if X_{2i} is the only omitted variable correlated with the treatment variable X_{1i} and

$$\mathbb{E}[X_2|X_1, W] = \mathbb{E}[X_2|W]$$

then the OLS estimator of β_1 in the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_i + \varepsilon_i$$

is a consistent estimator for the causal effect of X_1 on Y (i.e. β_1).

Control Variables

- What would happen if we were observe both X_{2i} and its control variable W_i (besides the outcome, Y_i , and the treatment variable of interest X_{1i})? In other terms, what are the properties of the OLS estimator of β_3 in the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 W_{2i} + \varepsilon$$

- We know that the true DGP for Y_i is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

and, therefore, unless $\text{cov}(W_{2i}, X_{3i}) \neq 0$, it will be true that the estimator of the coefficient on W_{2i} in a regression that also includes X_{1i} and X_{2i} will converge to 0. The reason is that, once we control for X_{1i} and X_{2i} , the random variable W_i is not correlated with Y_i (i.e. it does not have a causal effect on Y_i -after controlling for X_{1i} and X_{2i} -, and it is not correlated with any other unobserved variable that has a causal effect on Y_i).

INCLUDED VARIABLE BIAS

Including Irrelevant Variables in the Regression

- The case of W_{2i} in the previous slide is a particular example of the more general “problem” of introducing irrelevant variables in a regression.
- Assume the following model is true:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \text{with } \mathbb{E}[\varepsilon | X_1] = 0.$$

and assume that we estimate the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

with

$$\mathbb{E}[\varepsilon | X_1, X_2] = 0.$$

- In this case, $\hat{\beta}_1$ converges in probability to β_1 and $\hat{\beta}_2$ converges to 0.
- However, adding the irrelevant regressor X_2 increases the variance of $\hat{\beta}_1$.

Including Irrelevant Variables in the Regression

- Adding an irrelevant covariate will always increase the variance of the estimator of the causal effect of interest. This might make you decide not to reject the null hypothesis that a treatment is irrelevant in cases in which, in fact, the causal effect of such treatment is different from zero.
- Besides, in cases in which the OLS estimator is biased, adding covariates might exacerbate the bias of our OLS estimator. This is the so-called *included variable bias*.

Included Variable Bias

- Suppose you have the following regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

where X_1 is the regressor of interest (i.e. we are interested in the parameter β_1). We assume that X_2 is not observed by the econometrician, but that $\beta_2 \neq 0$ and $\text{cov}(X_1, X_2) \neq 0$. We also assume that

$$\mathbb{E}[\varepsilon | X_1, X_2] = 0.$$

- As we know from previous slides, if we compute the OLS estimator of β_1 in the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

then, $\hat{\beta}_1$ is affected by omitted variable bias:

$$\text{plim}(\hat{\beta}_1) = \frac{\text{cov}(Y_i, X_{1i})}{\text{var}(X_{1i})} = \beta_1 + \beta_2 \frac{\text{cov}(X_{1i}, X_{2i})}{\text{var}(X_{1i})}.$$

Included Variable Bias

- Imagine that we observe a third variable, X_3 such that
 - $\text{cov}(X_3, X_1) \neq 0$,
 - $\text{cov}(X_3, X_2) = \text{cov}(X_3, \varepsilon) = 0$.
- Note that the second bullet point is something that the econometrician cannot test: we do not have data on X_2 and/or ε . Assume that, mistakenly, we wrongly assume that $\text{cov}(X_3, X_2) \neq 0$ and, therefore, we decide to introduce X_3 as a control variable for X_2 .
- Is there “any cost” in including the additional variable X_3 ?
- In other words, if we start from a regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_{1i}] \neq 0,$$

is there any cost of introducing a variable X_{3i} such that $\text{cov}(X_{1i}, X_{3i}) \neq 0$ and $\text{cov}(\varepsilon_i, X_{3i}) = 0$?

Included Variable Bias

- In order to answer this question, we first characterize the $plim$ of $\hat{\beta}_1$ in the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \varepsilon_i$$

- Using the Frisch-Waugh-Lovell Theorem, we know that

$$plim(\hat{\beta}_1) = \frac{cov(Y - \gamma_0 - \gamma_1 X_3, X_1 - \delta_0 - \delta_1 X_3)}{var(X_1 - \delta_0 - \delta_1 X_3)}$$

where γ_0 and γ_1 are the population regression coefficients you get from regressing Y on X_3 , and δ_0 and δ_1 are the population regression coefficients from regressing X_1 on X_3 (i.e. $\delta_1 = cov(X_1, X_3)/var(X_3)$)

- Doing some algebraic manipulations, one can show that

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(X_{1i}, X_{2i})}{var(X_1 - \delta_0 - \delta_1 X_3)}.$$

Included Variable Bias

- Is the bias of $\hat{\beta}_1$ smaller in the case in which we add X_3 as a regressor?
- In order to answer this question, we should compare the bias of $\hat{\beta}_1$ in the regression of Y on X_1

$$\beta_2 \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)},$$

to the bias of $\hat{\beta}_1$ in the regression of Y on both X_1 and X_3

$$\beta_2 \frac{\text{cov}(X_{1i}, X_{2i})}{\text{var}(X_1 - \delta_0 - \delta_1 X_3)}.$$

- Therefore, the answer to this question depends on whether $\text{var}(X_1)$ is smaller or larger than $\text{var}(X_1 - \delta_0 - \delta_1 X_3)$.

$$\begin{aligned}\text{var}(X_1 - \delta_0 - \delta_1 X_3) &= \text{var}(X_1) + \delta_1^2 \text{var}(X_3) - 2\delta_1 \text{cov}(X_1, X_3) \\ &= \text{var}(X_1) + \frac{(\text{cov}(X_1, X_3))^2}{\text{var}(X_3)} - 2 \frac{(\text{cov}(X_1, X_3))^2}{\text{var}(X_3)}\end{aligned}$$

Included Variable Bias

- Therefore,

$$\begin{aligned} \text{var}(X_1 - \delta_0 - \delta_1 X_3) &= \text{var}(X_1) - \frac{(\text{cov}(X_1, X_3))^2}{\text{var}(X_3)} \\ &\leq \text{var}(X_1) \end{aligned}$$

and

$$\beta_2 \frac{\text{cov}(X_{1i}, X_{2i})}{\text{var}(X_1 - \delta_0 - \delta_1 X_3)} \geq \beta_2 \frac{\text{cov}(X_{1i}, X_{2i})}{\text{var}(X_1)}.$$

- The bias of the OLS estimator of β_1 is actually bigger (in absolute value) after including the additional regressor X_3 !
- It is **not** true that including more variables is **always** better. It **might** reduce bias (e.g. suppose that X_3 is literally equal to X_2), but it might make it worse (e.g. X_3 is correlated with X_1 and not correlated with X_2).

MEASURES OF FIT

Measures of Fit

- Up until now, we have focused on studying the properties of the OLS estimator $\hat{\beta}$ as estimator of the parameter β that captures the causal effect of some treatment variable X on some outcome variable Y .
- However, the estimated linear regression is also the *best* predictor Y given X in the sample:

$$(\hat{\beta}_0, \hat{\beta}_1) = \operatorname{argmin}_{b_0, b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2.$$

- Given this property of the OLS estimated regression line, we might wonder how well the regression line describes or fits the data?
- It is very important to remember that **the answer to this question is irrelevant if** the only thing **we care about** is finding a good estimator of **the causal effect** of a treatment variable on an outcome variable.
- On the contrary, **the answer to this question is key if we want to** use a vector of covariate X to **predict** a variable Y .

Measures of Fit

- A measure of fit quantifies how well the OLS estimate of the multiple regression line describes or captures the variation in Y in the data.
- Here we consider three different measures of fit:
 - Standard error of the regression: SER.
 - R^2 or R-squared.
 - \bar{R}^2 or adjusted R-squared.

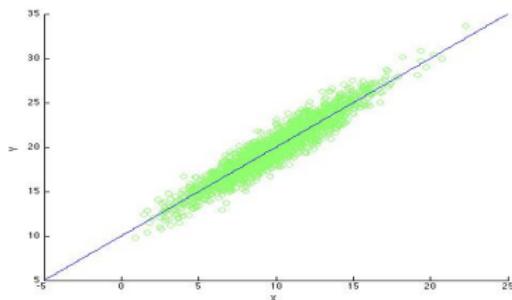


Figure : Good Fit

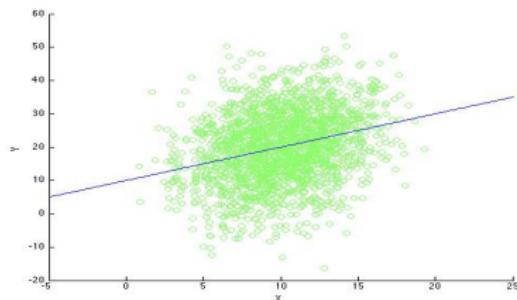


Figure : Bad Fit

In both figures, the regression line is exactly the same!

Standard Error of the Regression

- The standard error of the regression (SER) estimates the standard deviation of the error term, ε_i .
- Mathematically, the SER is

$$SER = \hat{\sigma}_\varepsilon, \quad \text{where } \hat{\sigma}_\varepsilon = \sqrt{\frac{1}{N - K - 1} \sum_{i=1}^N \hat{\varepsilon}_i^2} = \sqrt{\frac{SSR}{N - K - 1}},$$

where SSR is the sum of squared residuals, $SSR = \sum_{i=1}^n \hat{\varepsilon}_i^2$ and K is the total number of covariates in the regression.

- If the 4 standard OLS assumptions hold, then SER will be a consistent estimator for the standard deviation of the joint impact on the outcome variable of interest of all its determinants that are not explicitly included in the regression:

$$SER \xrightarrow{N \rightarrow \infty} \mathbb{V}(\varepsilon).$$

- In the left figure in slide 4, $SER = 1$. In the right figure in slide 4, $SER = 4$.

- The main problems of the SER as a measure of fit is that: (1) it is defined in the same units as the dependent variable, Y ; (2) any given SER gives information about the explanatory power of a regression only when compared with the standard deviation of the dependent variable, Y .
- The R^2 does not suffer from any of these problems.
- The R^2 is the fraction of the **sample** variance of Y ; explained by the regressors. Equivalently, the R^2 is 1 minus the fraction of the **sample** variance of Y ; that is explained by the joint impact on the outcome variable of interest of all its determinants that are not explicitly included in the regression.

- The R^2 is defined as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$$

where

$$ESS = \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^N (\hat{\beta}X_i - \bar{Y})^2$$

$$SSR = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^N (Y_i - \hat{\beta}X_i)^2$$

$$TSS = \sum_{i=1}^N (Y_i - \bar{Y})^2$$

- Proof of the equivalence of both expression for the R^2 .
- Decompose the sample variance of Y into *explained* and *unexplained* parts

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\&= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \\&= \sum_{i=1}^n \hat{\varepsilon}_i^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n \varepsilon_i(\hat{Y}_i - \bar{Y}) \\&= \sum_{i=1}^n \hat{\varepsilon}_i^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2 \sum_{i=1}^n \varepsilon_i(\hat{\beta}_1(X_i - \bar{X}_i) - \bar{\varepsilon}) \\&= \sum_{i=1}^n \hat{\varepsilon}_i^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2\hat{\beta}_1 \sum_{i=1}^n \varepsilon_i(X_i - \bar{X}_i) - \bar{\varepsilon} \sum_{i=1}^n \varepsilon_i \\&= \sum_{i=1}^n \hat{\varepsilon}_i^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + 2\hat{\beta}_1 0 - 0 \times 0\end{aligned}$$

- Proof of the equivalence of both expression for the R^2 (cont.)

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$TSS = SSR + ESS.$$

- Note that

$$\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^n (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X}) = 0 - \hat{\beta}_1 0 = 0$$

$$\sum_{i=1}^n \varepsilon_i (X_i - \bar{X}) = \sum_{i=1}^n [(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X})] (X_i - \bar{X})$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})^2$$

$$= \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X})^2 = 0.$$

Measures of Fit

- The R^2 ranges between 0 and 1, and its value does not depend on the units on which the random variables X and Y are measured.
- $R^2 = 1$ corresponds to a case in which $\hat{Y}_i = Y_i$ (i.e. the regression line $\hat{\beta}_0 + \hat{\beta}_1 X_i = \hat{Y}_i$ perfectly fits the observed data, Y_i).
- $R^2 = 0$ corresponds to a case in which $\hat{Y}_i = \bar{Y}$ (i.e. the regression line $\hat{\beta}_0 + \hat{\beta}_1 X_i$ does not vary with X_i)

Measures of Fit

OLS regression: STATA output

```
regress testscr str, robust
```

Regression with robust standard errors

Number of obs = 420
F(1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581

	Robust					
testscr	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-2.279808	.5194892	-4.39	0.000	-3.300945	-1.258671
_cons	698.933	10.36436	67.44	0.000	678.5602	719.3057

\bar{R}^2 or Adjusted R^2

- The R^2 is a measure of fit **in the sample**. Even if the 4 standard OLS assumptions hold, the R^2 is NOT a good measure of the fraction of population variance of Y that is explained by the term βX_i (i.e. by the composite effect of all the variables included in the regression).
- The reason is that the R^2 increases whenever a regressor is added, unless the OLS estimated coefficient on the added regressor is exactly 0.
- Even if we add covariates whose causal effect on Y is equal to 0 (i.e. we add covariates X_k such that $\beta_k = 0$), in any finite sample, the OLS estimates of the coefficients on these irrelevant covariates will generally be different from 0. Therefore, even if we add to our regression covariates whose causal effect is 0, the R^2 will generally increase.
- The adjusted R^2 or \bar{R}^2 modifies the R^2 so that it does not necessarily increase when a new regressor is added.

\overline{R}^2 or Adjusted R^2

- The \overline{R}^2 is

$$\overline{R}^2 = 1 - \frac{N-1}{N-K-1} \frac{SSR}{TSS} = 1 - \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_Y^2}$$

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N-K-1} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

$$\hat{\sigma}_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$$

- The \overline{R}^2 is always smaller than R^2 (i.e. $N-1 > N-K-1$ always)
- When we add a new covariate, the \overline{R}^2 might increase (because SSR becomes smaller) or decrease (because $N-K-1$ becomes smaller).

Warning Against Misinterpreting the R^2 or \bar{R}^2

- A high R^2 or \bar{R}^2 is NOT an indication that the OLS estimator $\hat{\beta}$ is a good estimator of the causal effect β .

A high R^2 or \bar{R}^2 only indicates the covariates included in the regression are very good predictors of Y , but it does not say anything about whether the relationship between X and Y is causal or not. Examples:

- If we regress the number of children living in a neighborhood on the number of snowmen built after a storm, the R^2 might be quite high...but this does not mean that building snowmen is a solution to the aging population.
- If we regress height on weight, the R^2 is very likely to be high...but this does not mean that getting fatter will make you taller.
- If we regress the number of Olympic gold medals in long-distance running on a dummy variable for wearing an Ethiopian jersey, the R^2 will be very high...but this does not mean that buying an Ethiopian jersey on Amazon will make me an Olympic gold medallist.

WWS 507c: Quantitative Analysis

Lecture 14: Which variables shall I include in my regression?

Princeton University

November 24, 2014

Summary of Previous Lectures

- Let's consider the following statistical model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \mathbb{E}[\varepsilon | X_1, X_2] = 0.$$

- The parameter of interest is the causal effect of X_1 on Y : β_1 .
- Let's consider 4 different cases depending on which data is available to you:
 - you observe $\{Y_i, X_{1i}; i = 1, \dots, N\}$
 - you observe $\{Y_i, X_{1i}, X_{2i}; i = 1, \dots, N\}$
 - you observe $\{Y_i, X_{1i}, X_{2i}, X_{3i}; i = 1, \dots, N\}$
 - you observe $\{Y_i, X_{1i}, X_{3i}; i = 1, \dots, N\}$
- For each of the four cases above, we will consider below two situations:
 - $\mathbb{E}[X_2 | X_1] = 0$, (this implies that $\text{cov}(X_1, X_2) = 0$)
 - $\mathbb{E}[X_2 | X_1] \neq 0$, (this is implied by $\text{cov}(X_1, X_2) \neq 0$)

Summary of Previous Lectures

Case 1: imagine you observe $\{Y_i, X_{1i}; i = 1, \dots, N\}$.

- The only regression you can run is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- If $\mathbb{E}[X_2|X_1] = 0$, then $plim(\hat{\beta}_1) = \beta_1$.
- If $\mathbb{E}[X_2|X_1] \neq 0$, then

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(X_1)}.$$

Summary of Previous Lectures

Case 2: imagine you observe $\{Y_i, X_{1i}, X_{2i}; i = 1, \dots, N\}$.

- You can choose between running two regressions. Either you ignore the data on $\{X_{2i}, i = 1, \dots, N\}$ and run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (1)$$

or you use the data on $\{X_{2i}, i = 1, \dots, N\}$ and run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (2)$$

- If $\mathbb{E}[X_2|X_1] = 0$, then, independently of whether you run regression (1) or regression (2), it will be true that

$$plim(\hat{\beta}_1) = \beta_1.$$

It is unclear whether the variance of $\hat{\beta}_1$ will be larger if you run regression (1) or if you run regression (2): it depends on $\text{cov}(X_1, X_2)$, β_2 and $\text{var}(X_2)$.

Summary of Previous Lectures

Case 2: imagine you observe $\{Y_i, X_{1i}, X_{2i}; i = 1, \dots, N\}$ (cont.)

- If $\mathbb{E}[X_2|X_1] \neq 0$ and you run regression (1), then

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(X_1)}.$$

If you include X_2 in your regression and run regression (2), then:

$$plim(\hat{\beta}_1) = \beta_1.$$

Summary of Previous Lectures

Case 2: imagine you observe $\{Y_i, X_{1i}, X_{2i}; i = 1, \dots, N\}$ (cont.)

- In summary, if you observe a variable X_2 that you think has a causal impact on your outcome variable of interest, Y , always add it as a regressor in your specification.
 - If that variable is correlated with your treatment variable of interest, you will avoid an omitted variable bias problem in your estimator.
 - If that variable is mean independent of your treatment variable of interest, the worst think that can happen is that the variance of your estimator of the causal effect of interest, β_1 , increases a little bit. This will happen only if the causal effect of X_2 on Y after controlling for X_1 , β_2 , is very small.

Summary of Previous Lectures

Case 3: imagine you observe $\{Y_i, X_{1i}, X_{2i}, X_{3i}; i = 1, \dots, N\}$.

- You can choose between running three regressions. Either you ignore the data on $\{X_{2i}, X_{3i}, i = 1, \dots, N\}$ and run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i \quad (3)$$

or you use the data on $\{X_{2i}, i = 1, \dots, N\}$ and ignore the data on $\{X_{3i}, i = 1, \dots, N\}$ and run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (4)$$

or you use all the available data and run the regression

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad (5)$$

- In Case 2, we already indicated that the regression in equation (4) is preferred over that in equation (3). Therefore, we focus here on the choice between including X_{3i} in the regression or not (i.e. eq. (5) vs. equation (4)).

Summary of Previous Lectures

Case 3: imagine you observe $\{Y_i, X_{1i}, X_{2i}, X_{3i}; i = 1, \dots, N\}$. (cont.)

- In slide 2 we had assumed that

$$\mathbb{E}[\varepsilon|X_1, X_2] = \mathbb{E}[Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2|X_1, X_2] = 0.$$

This implies that we are in one of the two following scenarios:

- Once we control for X_1 and X_2 , X_3 does not have a causal impact on Y (the term ε does not include any function of X_3). In this case, there is no reason to include X_3 in the regression. It will just increase the variance of our estimator of β_1 .
 - Once we control for X_1 and X_2 , X_3 has a causal impact on Y (the term ε includes functions of X_3), but X_3 is mean independent of both X_1 and X_2 . In this case, including X_3 will generally decrease the variance of our estimator of β_1 . Therefore, it is a good idea to also include X_3 .
- In summary, including X_3 or not never affects the consistency of the OLS estimator $\hat{\beta}_1$ as an estimator of β_1 .

Summary of Previous Lectures

Case 4: imagine you observe $\{Y_i, X_{1i}, X_{3i}; i = 1, \dots, N\}$.

- You can choose between running two regressions. Either you ignore the data on $\{X_{3i}, i = 1, \dots, N\}$ and run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i, \quad (6)$$

or you use all the available data and run

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{3i} + \varepsilon_i. \quad (7)$$

- If you run the regression in equation (6), then we are in the same situation as in case 1:

if $\mathbb{E}[X_2|X_1] = 0$, then $plim(\hat{\beta}_1) = \beta_1$,

if $\mathbb{E}[X_2|X_1] \neq 0$, then $plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(X_1)}$.

Summary of Previous Lectures

Case 4: imagine you observe $\{Y_i, X_{1i}, X_{3i}; i = 1, \dots, N\}$. (cont.)

- If $\mathbb{E}[X_2|X_1] = 0$, then, independently of whether we run the regression in equation (6) or that in equation (7), it will be true that $\hat{\beta}_1$ is a consistent estimator of β_1 .

Therefore, if we are willing to assume that $\mathbb{E}[X_2|X_1] = 0$, then the decision between equation (6) and equation (7) will only affect the variance of $\hat{\beta}_1$. This variance will be larger in equation (6) or in equation (7) depending on the joint distribution of (X_1, X_2, X_3) .

Summary of Previous Lectures

Case 4: imagine you observe $\{Y_i, X_{1i}, X_{3i}; i = 1, \dots, N\}$. (cont.)

- If $\mathbb{E}[X_2|X_1] \neq 0$, then the asymptotic bias of $\hat{\beta}_1$ as an estimator of β_1 might be larger (or smaller) in equation (7) than in equation (6).
- If X_3 is a good control for the missing variable X_2 ; i.e.

$$\mathbb{E}[X_2|X_3, X_1] = \mathbb{E}[X_2|X_3]$$

then the estimator $\hat{\beta}_1$ in the regression in equation (7) converges in probability to β_1 . In this case, we clearly prefer to include X_3 as a regressor.

Intuitively, you should think that X_3 is more likely to be a good control for X_2 whenever $\text{corr}(X_2, X_3)$ is high in absolute value (i.e. whenever $\text{corr}(X_2, X_3)$ is close to either 1 or -1).

Summary of Previous Lectures

Case 4: imagine you observe $\{Y_i, X_{1i}, X_{3i}; i = 1, \dots, N\}$. (cont.)

- If X_3 is not correlated with the missing variable X_2 , nor with the treatment variable of interest X_1 , then

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(X_1)},$$

Note that, in this case, the OLS estimator $\hat{\beta}_1$ converges to the same number independently of whether we include X_3 in the regression or not.

- If X_3 is not correlated with the missing variable X_2 but it is correlated with the treatment variable of interest X_1 , then

$$plim(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{cov(X_1, X_2)}{var(\tilde{X}_1)}, \quad \text{with } var(\tilde{X}_1) < var(X_1).$$

In this case, the OLS estimator $\hat{\beta}_1$ is larger in absolute value in equation (7) than in equation (6): we would prefer NOT to include X_3 in the regression. This is the so called *included variable bias*.

Summary of Previous Lectures

- In summary:

Q. Which variables shall I include in a regression?

A. Very broadly, these are the rules you should follow

- ① *Always include the treatment variable of interest.*
- ② *Always include any other variable that, after conditioning on the treatment variable of interest, has a causal effect on the outcome variable of interest.*
- ③ *Never include variables that, after conditioning on the treatment variable of interest (or other control variables you have already included), do not have an additional causal effect on the outcome variable of interest.*

WWS 507c: Quantitative Analysis

Lecture 15: Hypothesis Tests-Confidence Intervals

Princeton University

November 24, 2014

Introduction

- In previous lectures, we tried to answer the following question:

If we have a random sample on an outcome variable, Y_i , a treatment variable, X_{1i} , and other additional variables or controls, X_{2i}, \dots, X_{Ki} , which specification or regression shall we run so that we can use these data in order to obtain a consistent estimator of the causal effect of X_{1i} on Y_i ?
- For this lecture, let's assume that we have already found an specification

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki} + \varepsilon_i,$$

such that

$$\mathbb{E}[\varepsilon_i | X_{1i}, \dots, X_{Ki}] = \mathbb{E}[Y_i - \beta_0 - \beta_1 X_{1i} - \dots - \beta_K X_{Ki} | X_{1i}, \dots, X_{Ki}] = 0.$$

Introduction

- In this lecture, we will learn to answer questions of the kind
 - ① Is β_1 equal to 0? (i.e. is the causal effect of X_1 on Y equal to 0?)
 - ② Can we find an interval $[\underline{\beta}_1, \bar{\beta}_1]$ such that β_1 is contained in that interval with a certain probability?
 - ③ Is β_1 equal to β_2 ? (i.e. is the causal effect of X_1 on Y identical to that of X_2 ?)
 - ④ Are both β_1 and β_2 equal to 0?
- In lecture 11, we had already answered questions 1 and 2 for the case in which our regression model only includes one covariate. Now we will see that our answer easily generalizes to the case with more than one covariate.
- Questions 3 and 4 are new.

HYPOTHESIS TEST

Hypothesis Test

- Given a model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \varepsilon_i,$$

with

$$\mathbb{E}[\varepsilon_i | X_{1i}, \dots, X_{Ki}] = \mathbb{E}[Y_i - \beta_0 - \beta_1 X_{1i} - \dots - \beta_K X_{Ki} | X_{1i}, \dots, X_{Ki}] = 0,$$

in the following slides, we will learn to test the three following types of assumptions:

$$H_0 : \beta_k = \beta_0$$

vs.

$$H_1 : \beta_k \neq \beta_0,$$

$$H_0 : \beta_k = \beta_{k'} = \beta_0$$

vs.

$$H_1 : \beta_k \neq \beta_0 \text{ and/or } \beta_{k'} \neq \beta_0,$$

$$H_0 : \beta_k = \beta_{k'}$$

vs.

$$H_1 : \beta_k \neq \beta_{k'}.$$

HYPOTHESIS TEST OF A SINGLE COEFFICIENT

Hypothesis Test for a Single Coefficient

- Let's imagine that you want to test whether a change in class size has any effect on test scores. Formally, you want to test

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0.$$

- The OLS estimator $\hat{\beta}_1$ verifies that, **under the null hypothesis**,

$$\frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} \xrightarrow{d} \mathbb{N}(0, 1).$$

- In any finite sample, the point estimate $\hat{\beta}_1$ will generally be different from 0. How can we use the information in our sample to reject (or not) the hypothesis that the causal effect of class size on tests scores is exactly equal to 0 in the population (i.e. that $\beta_1 = 0$ in the population)? We can compute the p-value of the test:

$$\text{p-value} = 2 \times \Phi(-|\hat{\beta}_1/\hat{\sigma}_{\hat{\beta}_1}|) = 2 \times \left(1 - \Phi(|\hat{\beta}_1/\hat{\sigma}_{\hat{\beta}_1}|)\right)$$

Hypothesis Test for a Single Coefficient

- Another alternative would be to compute the a confidence interval that is known to contain the true value of β_1 :

95% confidence interval for β_1 =

$$(\hat{\beta}_1 - 1.96 \times \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + 1.96 \times \hat{\sigma}_{\hat{\beta}_1})$$

- STATA provides information on t-statistic, p-value, and 95% CI:

OLS regression: STATA output

```
regress testscr str, robust

Regression with robust standard errors
Number of obs = 420
F( 1, 418) = 19.26
Prob > F = 0.0000
R-squared = 0.0512
Root MSE = 18.581

-----
|           Robust
testscr |      Coef.  Std. Err.      t     P>|t|    [95% Conf. Interval]
-----+
str |  -2.279808  .5194892    -4.39    0.000    -3.300945   -1.258671
_cons |  698.933   10.36436    67.44    0.000    678.5602    719.3057
-----+
```

JOINT HYPOTHESIS TEST

Joint Hypothesis Test

- Imagine that you want to test that neither class size nor average income per capita affects test grades:

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0.$$

- This is just a particular case of the test

$$H_0 : \beta_1 = \beta_{10}, \beta_2 = \beta_{20} \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_{10} \text{ and/or } \beta_2 \neq \beta_{20},$$

for a particular pair of values (β_{10}, β_{20}) .

- This null hypothesis imposes two restrictions on the multiple regression model: it is a joint hypothesis.
- If one (or more than one) of the equalities under the null hypothesis is false, then the joint hypothesis itself is false. The alternative hypothesis is that at least one of the equalities in the null hypothesis is false.

Joint Hypothesis Test

- How can we test two restrictions at the same time?

Alternative 1:

- You could be tempted to test H_0 through the following procedure:
 - ① First, test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.
 - ② Second, test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$.
 - ③ Third, reject the null $H_0 : \beta_1 = \beta_2 = 0$ at a significance level α if and only if you have rejected either $H_0 : \beta_1 = 0$ or $H_0 : \beta_2 = 0$ at a significance level α .
- This procedure is NOT correct: it makes you reject $H_0 : \beta_1 = \beta_2 = 0$ more than $\alpha\%$ of the times!. The reason is that this procedure does not take into account the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$.
- Intuition. Based on this procedure, $P(\{\text{not rejecting } H_0 : \beta_1 = \beta_2 = 0\})$ is:

$$P(\{\text{not reject } H_0 : \beta_1 = 0\} \cap \{\text{not reject } H_0 : \beta_2 = 0\}).$$

Joint Hypothesis Test

- How can we test two restrictions at the same time?

Alternative 1: (cont.)

- Intuition (cont.). Imagine that $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent, then

$$P(\{\text{not reject } H_0 : \beta_1 = 0\} \cap \{\text{not reject } H_0 : \beta_2 = 0\}) =$$

$$P(\{\text{not reject } H_0 : \beta_1 = 0\}) \times P(\{\text{not reject } H_0 : \beta_2 = 0\}) =$$

$$(1 - \alpha) \times (1 - \alpha)$$

and note that $(1 - \alpha)^2 < (1 - \alpha)$. Therefore, this alternative 1 procedure tends to reject the null hypothesis

$$H_0 : \beta_1 = \beta_2 = 0$$

too often.

Joint Hypothesis Test

- How can we test two restrictions at the same time?

Alternative 2:

- Perform the so-called **F-test**. There are two versions of the F-test:
 - heteroskedasticity-robust.
 - homoskedasticity-only.

The heteroskedasticity-robust F-test is preferred. In order to compute it in STATA, after

```
reg y x1 x2 x3, r
```

you should type

```
test x1 x2
```

STATA will give you the p-value of the test.

Joint Hypothesis Test

- How can we test two restrictions at the same time?

Alternative 2: (cont.)

- With the exclusive purpose of gaining intuition about how the F-test works, we will focus our attention on the homoskedasticity-only F-statistic. For expositional purposes, the advantage of the homoskedasticity-only F-statistic is that it can be written as a function of familiar objects (e.g. SSR and R^2).
- Let's imagine that we are trying to test

$$H_0 : \beta_1 = \beta_2 = 0 \text{ vs. } H_1 : \beta_1 \neq 0, \text{ and/or } \beta_2 \neq 0$$

in a model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i.$$

This is called the **unrestricted model**. Once we impose the restrictions included in the null hypothesis, we obtain the so-called **restricted model**:

$$Y_i = \beta_0 + \beta_3 X_{3i} + \varepsilon_i.$$

Joint Hypothesis Test

- How can we test two restrictions at the same time?

Alternative 2: (cont.)

- If the null hypothesis was true, in a random sample from the population of interest, we would expect the fit of the unrestricted model to be similar to the fit of the restricted model. If the null hypothesis was not true, then the unrestricted model should fit the data significantly better than the restricted model.
- Using this intuition, the F-statistic used in the homoskedasticity-only F-test is:

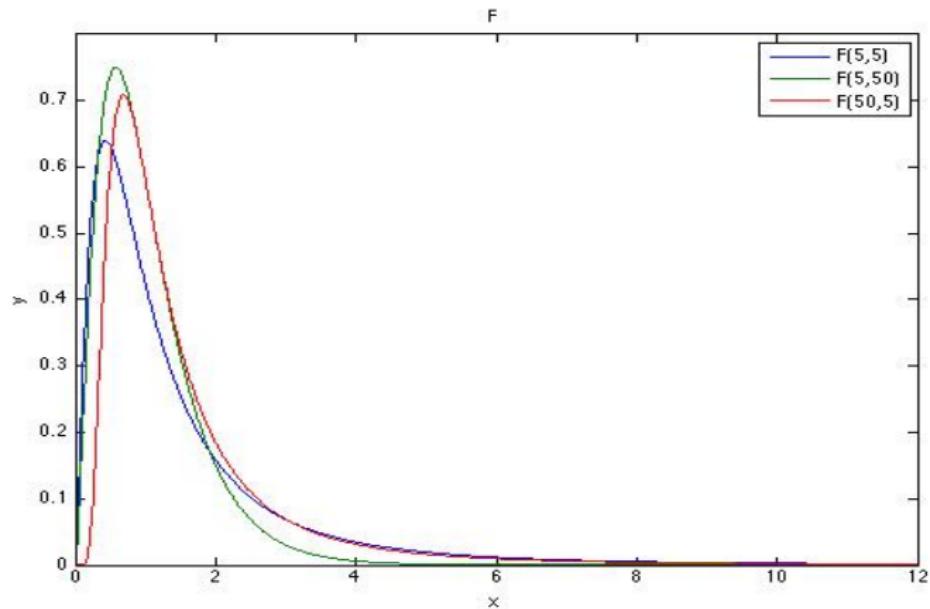
$$F = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)},$$

where q is the number of restrictions we are simultaneously testing.

- Under homoskedasticity, this F-statistic is distributed as an F-distribution with degrees of freedom (q, ∞) .

F Distribution

- How does the F distribution look like?



TEST OF A SINGLE RESTRICTION INVOLVING MULTIPLE PARAMETERS

Test of Single Restriction Involving Multiple Parameters

- Imagine that you want to test whether β_1 is equal to β_2 in a model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

i.e. $H_0 : \beta_1 = \beta_2$ vs. $H_1 : \beta_1 \neq \beta_2$

- There are two alternative ways to test this restriction using STATA,

- Write

```
reg Y X1 X2, r
```

```
test X1=X2
```

- Note that testing the null hypothesis $H_0 : \beta_1 = \beta_2$ is equivalent to testing

$$H_0 : \beta_1 - \beta_2 = 0,$$

and transform the regression so that the coefficient on one of the covariates of the transformed regression is equal to $\beta_1 - \beta_2$. Then use the t-statistic to test whether this coefficient is equal to 0 or not.

Test of Single Restriction Involving Multiple Parameters

- As an example of how to transform a regression, imagine that you want to test whether β_1 is equal to β_2 in a model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad \mathbb{E}[\varepsilon | X_{1i}, X_{2i}] = 0 \quad (1)$$

Adding and subtracting $\beta_2 X_{1i}$, we obtain:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_2 X_{1i} - \beta_2 X_{1i} + \varepsilon_i, \\ Y_i &= \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + \varepsilon_i, \\ Y_i &= \beta_0 + \gamma X_{1i} + \beta_2 X_{3i} + \varepsilon_i, \end{aligned} \quad (2)$$

with $\gamma = \beta_1 - \beta_2$ and $X_{3i} = X_{1i} + X_{2i}$. Therefore, if model (1) is right, testing $H_0 : \beta_1 - \beta_2 = 0$ in equation (1) is equivalent to testing $H_0 : \gamma = 0$ in equation (2). Given that

$$\frac{\hat{\gamma} - (\beta_1 - \beta_2)}{SE(\hat{\beta}_1 - \hat{\beta}_2)} = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \xrightarrow{d} \mathbb{N}(0, 1).$$

we can use a t test to perform this test.

CONFIDENCE SETS FOR MULTIPLE COEFFICIENTS

Confidence Sets for Multiple Coefficients

- In order to get intuition about how to build confidence sets for multiple coefficients, let's think again about the logic behind the confidence intervals for a single coefficient.
- In the case in which we want to perform inference on a single coefficient (e.g. β_1), there is a very close connection between the statistic we use to perform tests,

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

and the exact expression that we use to form confidence intervals,

$$[\hat{\beta}_1 - Z_{1-\alpha/2} \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + Z_{1-\alpha/2} \hat{\sigma}_{\hat{\beta}_1}]$$

In particular, we form confidence intervals by *inverting* the test statistic; i.e. in a confidence interval with confidence level $1 - \alpha$, we accept all those values of β_1 that we would not reject in a test with significance level α .

Confidence Sets for Multiple Coefficients

- In the case of multiple coefficients, we can also build confidence intervals by *inverting* a test statistic. In particular, we saw above that we can use an F-test to test the joint null hypothesis:

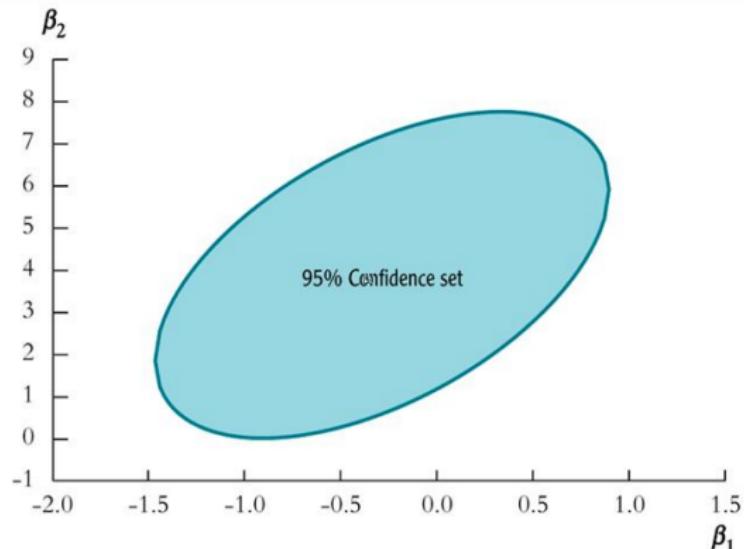
$$H_0 : \beta_1 = \beta_{10}$$

$$\beta_2 = \beta_{20}$$

versus the alternative that at least one of these equalities is not true. One way of building a 95% confidence interval for the pair (β_1, β_2) is to try all the possible pairs (β_{10}, β_{20}) , perform an F-test on each of them, and keep in the confidence interval only those particular values of the pair (β_{10}, β_{20}) that are not rejected at the 5% level.

Confidence Sets for Multiple Coefficients

FIGURE 5.1 95% Confidence Set for β_1 and β_2



The 95% confidence set for β_1 and β_2 is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the F -statistic at the 5% significance level.

WWS 507c: Quantitative Analysis

Lecture 16: Nonlinear Regression Functions

Princeton University

November 24, 2014

Introduction

- All the regression models that we have seen so far are of the type

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_K X_{Ki} + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_{1i}, \dots, X_{Ki}] = 0.$$

- All these models are linear in the covariates X_{1i}, \dots, X_{Ki} .
- This linearity implies that the effect of increasing some covariate X_k in one unit on the conditional expectation

$$\mathbb{E}[Y | X_1, \dots, X_K]$$

is constant; i.e. (a) independent of the initial value of X_k ; (b) independent of the value of any other causal determinants of Y (i.e. independent of $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_K$).

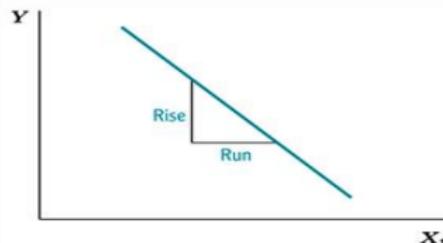
- Mathematically:

$$\mathbb{E}[Y | X_1 + \Delta, \dots, X_K] - \mathbb{E}[Y | X_1, \dots, X_K] = \beta_1 \Delta,$$

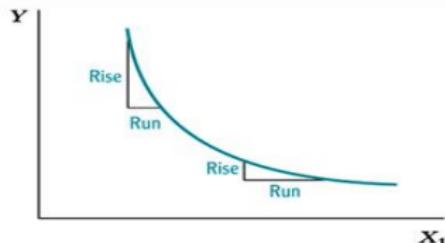
where Δ is any real number.

Introduction

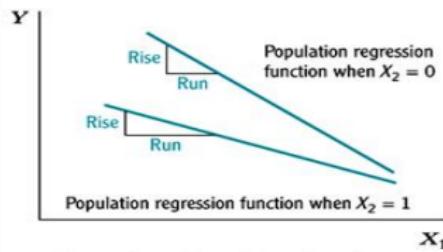
FIGURE 6.1 Population Regression Functions with Different Slopes



(a) Constant slope



(b) Slope depends on the value of X_1



(c) Slope depends on the value of X_2

In Figure 6.1a, the population regression function has a constant slope. In Figure 6.1b, the slope of the population regression function depends on the value of X_1 . In Figure 6.1c, the slope of the population regression function depends on the value of X_2 .

Introduction

- In many cases of interest, causal effects of treatment on outcomes are nonlinear. E.g.
 - Do we think that the effect on grades of increasing class size from 5 to 10 students will be same as increasing it from 70 to 75 students?
 - Do we think that the effect on grades of decreasing class size from 30 to 25 students will depend on the average temperature and location of the school?
 - In the sunny beach town of (.....), students don't go to class in any case, so it does not matter much how many students are enrolled in each class.
- In this lecture, we will learn how to write two types of nonlinear models:
 - models in which the effect on Y of a change in one unit in a variable X_k depends on the value of X_k itself.
 - models in which the effect on Y of a change in one unit in a variable X_k depends on the value of another independent variable $X_{k'}$.
- In all models we will study in this lecture, it will still be true that:
 - the only unobservable term in the regression enters linearly (i.e. ε is additive)
 - the regression function is still linear in the parameters (i.e. the regression function is linear in β_k , for every k)

Introduction

- Example of models that are nonlinear in covariates but linear in parameters and with additive error term.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

$$\ln(Y_i) = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

In order to estimate $(\beta_0, \beta_1, \beta_2)$, we will still use OLS

- Example of a model with a non-additive error term:

$$Y_i = \beta_0 + \beta_1 (X_i)^{\varepsilon_i}$$

In order to estimate $(\beta_0, \beta_1, \beta_2)$, we will NOT use OLS

- Example of a model with additive error term but nonlinear in both parameters and covariates:

$$Y_i = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_i)} + \varepsilon_i$$

In order to estimate $(\beta_0, \beta_1, \beta_2)$, we will NOT use OLS

CAUSAL EFFECT OF X_k DEPENDS ON THE VALUE OF X_k

Causal Effect of X_k Depends on the Value of X_k

- There are two general models that allow for the causal effect of a change in a variable X_k to depend on the initial value of X_k :

- Polynomial regression model of degree r :

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_r X_{1i}^r + \varepsilon_i$$

- Logarithmic models. We will study three different models:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i$$

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i$$

- For all these models, the OLS estimator of the parameter vector will be well-defined, unbiased, consistent and asymptotically normal.

POLYNOMIAL REGRESSION MODELS

Polynomial Regression Models

- The polynomial regression model is a specific case of the multiple regression model we saw in Lectures 13 to 15. The only difference is that, in Lectures 13 to 15, the regressors were different independent variables, while here the regressors are powers of the same independent variable X .

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_r X_{ri} + \varepsilon_i$$

vs.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \cdots + \beta_r X_{1i}^r + \varepsilon_i$$

- The techniques for estimation and inference that we saw in Lectures 13 to 15 also apply here. In particular, we can test the null hypothesis that the relationship between Y and X_1 is linear, versus the alternative hypothesis that it is nonlinear, by using an F-test to test for,

$$H_0 : \beta_2 = \cdots = \beta_r = 0, \quad \text{vs.} \quad H_0 : \text{at least one } \beta_j \neq 0, j = 2, \dots, r.$$

Polynomial Regression Models

- What happens if the correct specification is a quadratic regression but we estimate a linear model?
- If the correct specification is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0,$$

but we estimate

$$Y_i = \gamma_0 + \gamma_1 X_i + u_i,$$

then

$$u_i = \beta_2 X_i^2 + \varepsilon_i,$$

and

$$\mathbb{E}[u_i | X_i] = \mathbb{E}[\beta_2 X_i^2 + \varepsilon_i | X_i] = \beta_2 X_i^2,$$

which will be different from 0 unless $\beta_2 = 0$.

Polynomial Regression Models

- Therefore, using the formula for the omitted variable bias, we know that the OLS estimator of γ_1 in the regression $Y = \gamma_0 + \gamma_1 X + u$, will converge to

$$\gamma_1 = \beta_1 + \beta_2 \frac{\text{cov}(X, X^2)}{\text{var}(X)}$$

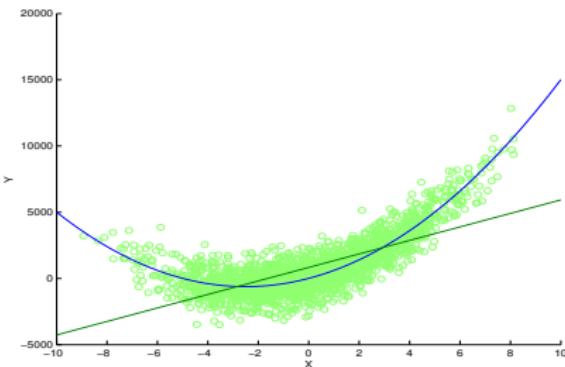


Figure: $\beta_2 > 0$

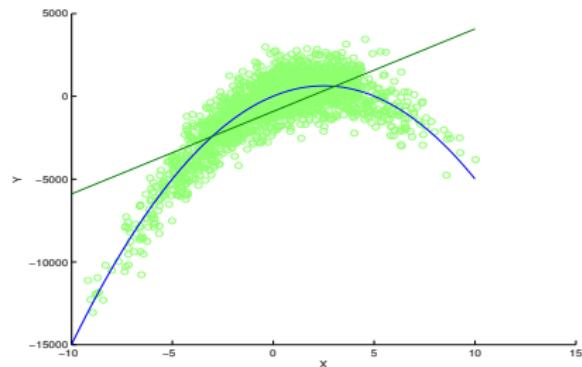


Figure: $\beta_2 < 0$

Polynomial Regression Models

- Does γ_1 have an easy interpretation?
- Imagine the correct specification is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0.$$

- In this model, the effect of increasing X_i by one unit is:

$$\beta_1 + \beta_2[(X_i + 1)^2 - X_i^2] = \beta_1 + \beta_2(2X_i + 1)$$

The causal effect of increasing X_i in one unit depends on the value of X_i .

- If we were to increase X_i in one unit *for every individual i in the population of interest*, then the **average causal effect** will be

$$\mathbb{E}[\beta_1 + \beta_2(2X_i + 1)] = \beta_1 + \beta_2(2\mathbb{E}[X_i] + 1)$$

- Note that

$$\beta_1 + \beta_2(2\mathbb{E}[X_i] + 1) \neq \beta_1 + \beta_2 \frac{\text{cov}(X, X^2)}{\text{var}(X)},$$

therefore, γ_1 is NOT the average causal effect.

LOGARITHMIC REGRESSION MODELS

Logarithmic Regression Models

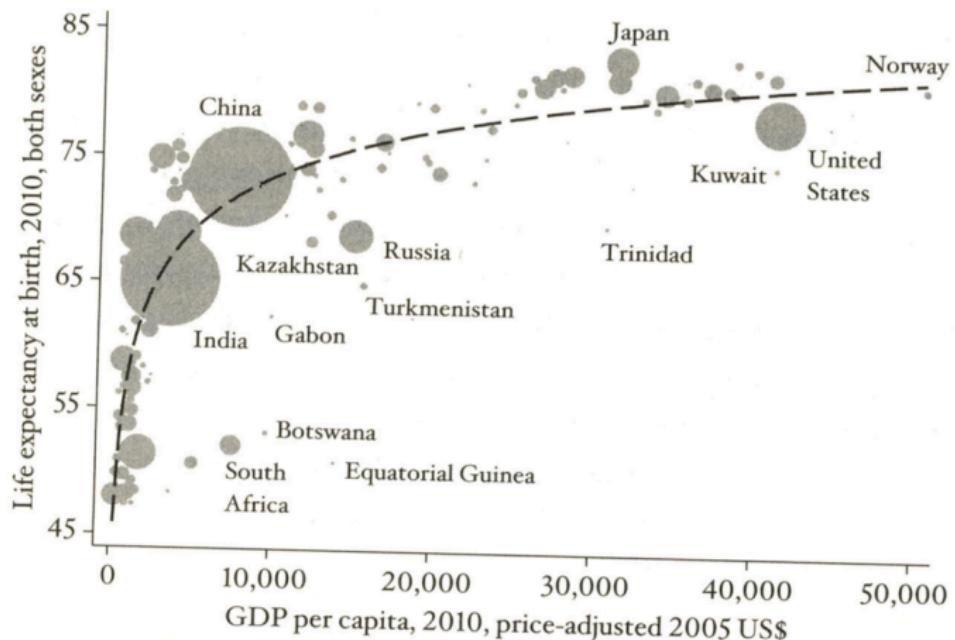


FIGURE 1 Life expectancy and GDP per capita in 2010.

Logarithmic Regression Models

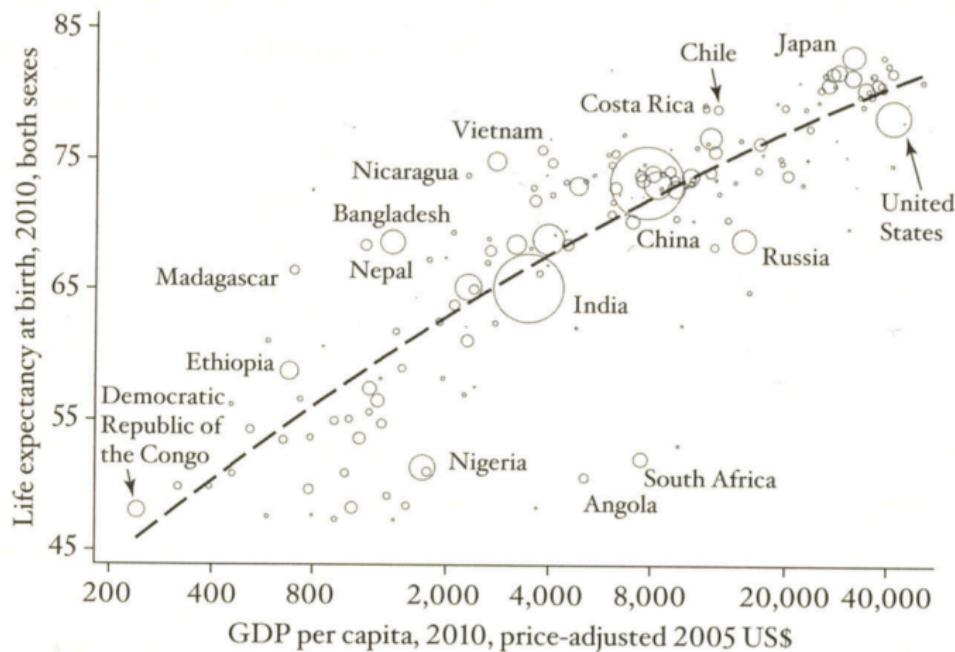


FIGURE 2 Life expectancy and GDP per capita in 2010 on a log scale.

Logarithmic Regression Models

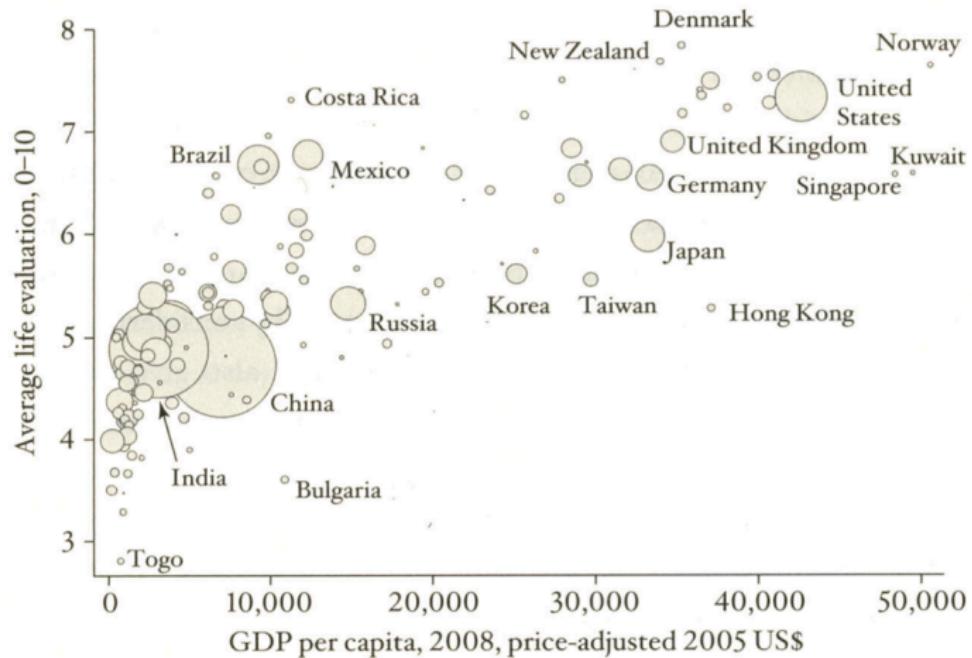


FIGURE 1 Life evaluation and GDP per capita.

Logarithmic Regression Models

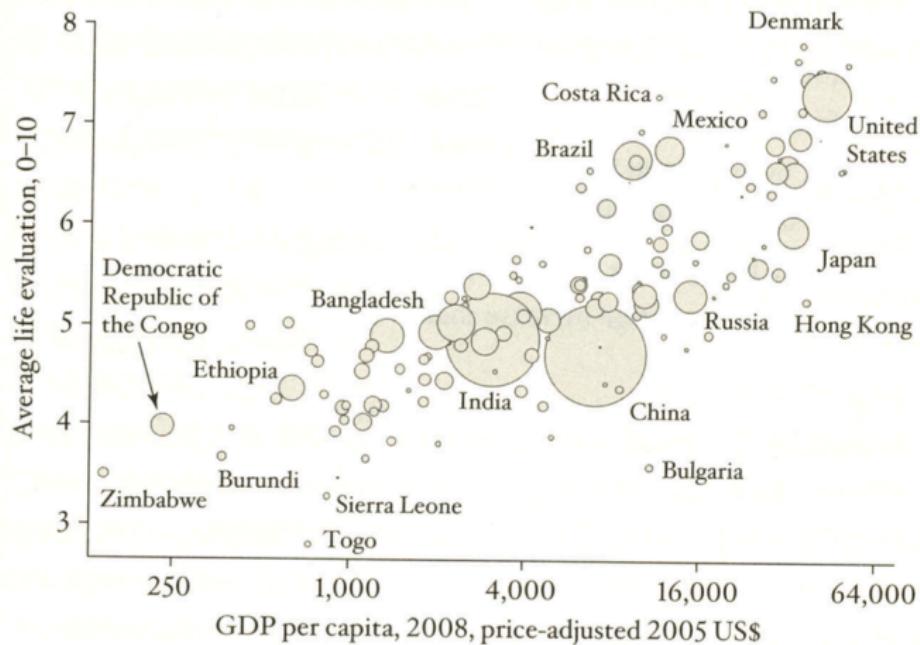


FIGURE 2 Life evaluation and GDP per capita on a log scale.

Logarithmic Regression Models

- Math review. If Z is small, then

$$\ln(1 + Z) \approx Z.$$

- Therefore, when $\Delta X/X$ (relative change in X) is small

$$\ln\left(1 + \frac{\Delta X}{X}\right) \approx \frac{\Delta X}{X}.$$

- Remember also that

$$\ln(A) - \ln(B) = \ln\left(\frac{A}{B}\right)$$

- Therefore,

$$\ln(X + \Delta X) - \ln(X) = \ln\left(\frac{X + \Delta X}{X}\right) = \ln\left(1 + \frac{\Delta X}{X}\right)$$

and, if $\Delta X/X$ is small,

$$\ln(X + \Delta X) - \ln(X) \approx \frac{\Delta X}{X}.$$

Logarithmic Regression Models

- Key relationship between changes in the natural logarithm of X and percentage changes in X :

When $\Delta X/X$ (relative change in X) is small, the difference between the logarithm of $X + \Delta X$ and the logarithm of X is approximately $\Delta X/X$.

- Therefore, when $\Delta X/X$ is small,

$$100(\ln(X + \Delta X) - \ln(X)) \approx 100\left(\frac{\Delta X}{X}\right),$$

where $100(\Delta X/X)$ is the percentage change in X .

- E.g. 1: imagine $X = 100$ and $\Delta X = 1$, then

$$100 \times (\ln(101) - \ln(100)) = 0.995 \text{ and } 100 \times \frac{1}{100} = 1.$$

- E.g. 2: imagine $X = 100$ and $\Delta X = 50$, then

$$100 \times (\ln(150) - \ln(100)) = 40.55 \text{ and } 100 \times \frac{50}{100} = 50.$$

Model 1: $Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i$, $\mathbb{E}[\varepsilon_i | X_i] = 0$

- How can we interpret β_1 ?
- Interpretation 1: Note that β_1 captures the causal effect on the expected value of Y of increasing $\ln(X)$ in one unit.
- Interpretation 1 is not very useful: we do not tend to think of policy evaluations that imply changes in natural logarithms of variables.
- Interpretation 2: if $\Delta X/X$ is small, then $\beta_1(\Delta X/X)$ is approximately the effect on the expected value of Y of increasing X in ΔX . Proof:

$$Y_i + \Delta Y - Y_i = \beta_0 + \beta_1 \ln(X + \Delta X) + \varepsilon_i - \beta_0 - \beta_1 \ln(X) - \varepsilon_i;$$

$$\begin{aligned}\Delta Y &= \beta_1(\ln(X + \Delta X) - \ln(X)) = \beta_1 \ln\left(1 + \frac{\Delta X}{X}\right) \\ &\approx \beta_1\left(\frac{\Delta X}{X}\right) = \frac{\beta_1}{100}\left(\frac{100\Delta X}{X}\%\right)\end{aligned}$$

- Therefore, if ΔX is small, $(\beta_1/100)$ is approximately the effect on the expected value on Y of increasing X in $100(\Delta X/X)\%$.

$$\text{Model 1: } Y_i = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0$$

- What happens if $\Delta X/X$ is large?
- Then the approximation

$$\ln\left(1 + \frac{\Delta X}{X}\right) \approx \frac{\Delta X}{X}$$

is a bad approximation, and we should compute the actual effect on Y individually for each value x of X :

$$\Delta Y = \beta_1 \ln\left(1 + \frac{\Delta X}{X}\right)$$

Model 2: $\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$, $\mathbb{E}[\varepsilon_i | X_i] = 0$

- How can we interpret β_1 ?
- Interpretation 1: Note that β_1 captures the causal effect on the expected value of $\ln(Y)$ of increasing X in one unit.

$$\mathbb{E}[\ln(Y)|X + \Delta X] - \mathbb{E}[\ln(Y)|X] = \beta_1 \Delta X.$$

- Interpretation 1 is not very useful: we do not tend to evaluate policy changes in terms of changes in the logarithm of our variable of interest.
- Interpretation 2: if $(\Delta Y/Y)$ is small, then $100\beta_1 \Delta X$ is approximately the percentage change in Y of changing X in ΔX . Proof:

$$\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$$

$$\ln\left(1 + \frac{\Delta Y}{Y}\right) = \beta_1 \Delta X$$

and, if $(\Delta Y/Y)$ is small, then

$$\ln\left(1 + \frac{\Delta Y}{Y}\right) \approx \frac{\Delta Y}{Y} \rightarrow 100 \frac{\Delta Y}{Y} \% \approx 100\beta_1 \Delta X$$

Model 2: $\ln(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$, $\mathbb{E}[\varepsilon_i | X_i] = 0$

- What happens if $\Delta X/X$ is large?
- Then, we can rewrite:

$$Y_i = \exp(\beta_0 + \beta_1 X_i + \varepsilon_i)$$

$$Y_i = \exp(\beta_0) \exp(\beta_1 X_i) \exp(\varepsilon_i)$$

and

$$\frac{Y + \Delta Y}{Y} = \frac{\exp(\beta_0) \exp(\beta_1 \Delta X) \exp(\beta_1 X_i) \exp(\varepsilon_i)}{\exp(\beta_0) \exp(\beta_1 X_i) \exp(\varepsilon_i)} = \exp(\beta_1 \Delta X)$$

or, equivalently,

$$\frac{\Delta Y}{Y} = \exp(\beta_1 \Delta X) - 1$$

and

$$100 \frac{\Delta Y}{Y} \% = 100(\exp(\beta_1 \Delta X) - 1)$$

$$\text{Model 3: } \ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0$$

- Interpretation 1: β_1 captures the causal effect on the expected value of $\ln(Y)$ of increasing $\ln(X)$ in one unit.

$$\mathbb{E}[\ln(Y) | \ln(X) + \Delta X] - \mathbb{E}[\ln(Y) | \ln(X)] = \beta_1 \Delta X.$$

- Interpretation 1 is not very useful: we do not tend to evaluate policy changes in terms of the effect of changes in the log of a treatment variable on the log of our outcome variable of interest.
- Interpretation 2: β_1 is the percentage change in Y of changing X in $100(\Delta X/X)\%$.

$$\ln(Y + \Delta Y) - \ln(Y) = \beta_1 (\ln(X + \Delta X) - \ln(X))$$

$$\ln\left(\frac{Y + \Delta Y}{Y}\right) = \beta_1 \ln\left(\frac{X + \Delta X}{X}\right)$$

$$\ln\left(1 + \frac{\Delta Y}{Y}\right) = \beta_1 \ln\left(1 + \frac{\Delta X}{X}\right)$$

and $\ln(1 + (\Delta Z/Z)) \approx (\Delta Z/Z) = 100(\Delta Z/Z)\%$, for $Z = X, Y$.

$$\text{Model 3: } \ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i | X_i] = 0$$

- What happens if $\Delta X/X$ or $\Delta Y/Y$ are large?
- Then we can rewrite:

$$Y = \exp(\beta_0) X^{\beta_1} \exp(\varepsilon)$$

$$Y + \Delta Y = \exp(\beta_0) (X + \Delta X)^{\beta_1} \exp(\varepsilon)$$

and

$$\frac{Y + \Delta Y}{Y} = \frac{\exp(\beta_0)(X + \Delta X)^{\beta_1} \exp(\varepsilon)}{\exp(\beta_0) X^{\beta_1} \exp(\varepsilon)}$$

$$\frac{Y + \Delta Y}{Y} = \frac{(X + \Delta X)^{\beta_1}}{X^{\beta_1}}$$

$$\frac{\Delta Y}{Y} = \frac{(X + \Delta X)^{\beta_1}}{X^{\beta_1}} - 1$$

$$100 \frac{\Delta Y}{Y} \% = 100 \left(\frac{(X + \Delta X)^{\beta_1}}{X^{\beta_1}} - 1 \right) \%$$

Additional Comments

- In Lecture 13, we discussed R^2 and \bar{R}^2 as measures of fit.
- It is important to bear in mind that both R^2 and \bar{R}^2 allow to compare the fit of different specifications, **as long as all these different specifications have the same dependent variable.**
- Therefore, we can use R^2/\bar{R}^2 to compare:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \text{vs.} \quad Y = \beta_0 + \beta_1 \ln(X) + \varepsilon,$$

and

$$\ln(Y) = \beta_0 + \beta_1 X + \varepsilon, \quad \text{vs.} \quad \ln(Y) = \beta_0 + \beta_1 \ln(X) + \varepsilon.$$

- Comparisons of R^2/\bar{R}^2 for regressions with different dependent variables do not make sense. R^2/\bar{R}^2 compare fractions of the variance of the dependent variable explained by different regressors. If the variance of the dependent variable changes across models, R^2/\bar{R}^2 cannot tell much about the relative fit of those models.

CAUSAL EFFECT OF X_k DEPENDS ON THE VALUE OF $X_{k'}$, FOR $k' \neq k$

Interactions Between Two Binary Variables

- In Spain there are two big political parties, PP and PSOE. Elections always take place on a Sunday. Let's define the following variables:
 - Y : Difference in number of votes between PSOE and PP.
 - D_F : dummy for "nice weather on Friday"
 - D_S : dummy for "nice weather on Sunday".
- We could write a regression as

$$Y = \beta_0 + \beta_1 D_F + \varepsilon,$$

with $\beta_1 > 0$.

- In this case,

$$\beta_1 = \mathbb{E}[Y|D_F = 1] - \mathbb{E}[Y|D_F = 0],$$

or, in words, the difference in the mean incumbent vote margin between those elections in which GDP went up in the last month before elections, and those elections in which GDP went down.

Interactions Between Two Binary Variables

- Accordingly, the OLS estimator in the previous regression is simply the difference in sample averages:

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0$$

with

$$\bar{Y}_1 = \frac{\sum_{i=1}^N Y_i D_{Fi}}{\sum_{i=1}^I D_{Fi}},$$
$$\bar{Y}_0 = \frac{\sum_{i=1}^N Y_i (1 - D_{Fi})}{\sum_{i=1}^I (1 - D_{Fi})}.$$

where we observe a random sample of N elections.

- Note that we can use a t-test on $\hat{\beta}_1$ in order to test whether the difference in sample means, $\bar{Y}_1 - \bar{Y}_0$, is statistically significant or not.

Interactions Between Two Binary Variables

- In the regression,

$$Y = \beta_0 + \beta_1 D_F + \varepsilon,$$

it is hard to believe that $\mathbb{E}[\varepsilon | D_F] = 0$. As an example of omitted variable, the weather on Sunday affects how many PSOE voters will actually vote and the weather on Sunday tends to be correlated with the weather on Friday.

Therefore, we could write

$$Y = \beta_0 + \beta_1 D_F + \beta_2 D_S + \varepsilon.$$

with $\beta_1 > 0$ and $\beta_2 < 0$.

- This regression assumes that the effect of having nice weather on Friday is independent of whether the weather is nice on Sunday. This is not true. If the weather is good on Friday but bad on Sunday, some of the voters that left for the weekend will come back. The number of PP voters that miss the election because they are on vacation will be larger if we have good weather on Friday **and** Sunday.

Interactions Between Two Binary Variables

- In order to capture the differential impact of the variable D_1 depending on the value of D_2 , we need to introduce an **interaction term**

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + \varepsilon,$$

with $\beta_1 > 0$, $\beta_2 < 0$, and $\beta_3 > 0$.

- Note that:

$$\mathbb{E}[Y|D_1 = 1, D_2 = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3,$$

$$\mathbb{E}[Y|D_1 = 1, D_2 = 0] = \beta_0 + \beta_1,$$

$$\mathbb{E}[Y|D_1 = 0, D_2 = 1] = \beta_0 + \beta_2,$$

$$\mathbb{E}[Y|D_1 = 0, D_2 = 0] = \beta_0,$$

- If we had not introduced the interaction term $(D_1 \times D_2)$, then

$$\mathbb{E}[Y|D_1 = 1, D_2 = 1] = \beta_0 + \beta_1 + \beta_2.$$

Interactions Between Discrete and Continuous Variable

- Consider the following variables
 - Y : wage
 - H : height
 - F : female dummy
- Interpret the following regressions:

$$Y = \beta_0 + \beta_1 H + \varepsilon$$

$$Y = \beta_0 + \beta_1 F + \varepsilon$$

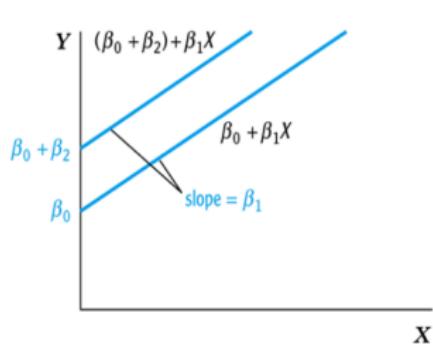
$$Y = \beta_0 + \beta_1 H + \beta_2 (H \times F) + \varepsilon$$

$$Y = \beta_0 + \beta_1 H + \beta_2 F + \varepsilon$$

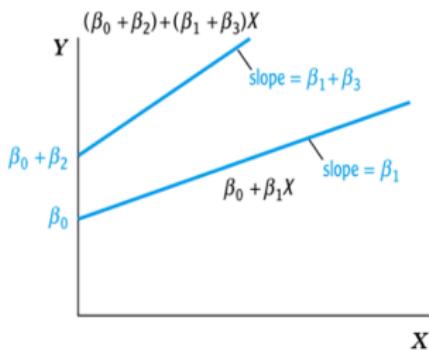
$$Y = \beta_0 + \beta_1 H + \beta_2 F + \beta_3 (H \times F) + \varepsilon$$

- Map the last three regressions in this list to the following figures:

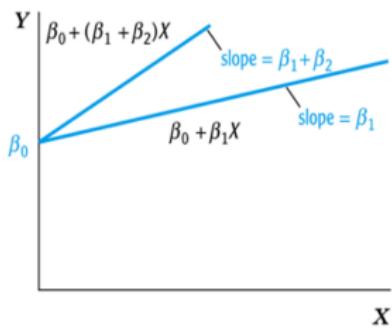
Interactions Between Discrete and Continuous Variable



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interactions Between Continuous Variables

- We can also write interactions effects between continuous variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon, \quad \mathbb{E}[\varepsilon | X_1, X_2] = 0$$

In this case, the causal effect of increasing X_1 in one unit is $\beta_1 + \beta_3 X_2$.

WWS 507c: Quantitative Analysis

Lecture 17

Princeton University

December 8, 2014

Naked Statistics

[http://www.npr.org/blogs/money/2013/04/23/178635250/
episode-453-what-causes-what](http://www.npr.org/blogs/money/2013/04/23/178635250/episode-453-what-causes-what)

- Research questions:
 - ① Effect of attending Chicago public schools on future performance;
 - ② Effect of hormone replacement therapy on likelihood of suffering from heart disease;
 - ③ Effect of police on crime;
 - ④ Effect of attending more selective colleges on earnings.

Effect of Police on Crime

- Do police deter crime? Answer: Klick and Tabarrok (2005)
- A majority of studies surveyed found that either there is no relationship or increases in the number of police are associated with increases in the level of crime.
- Most economists are suspicious of these results.
- It is no surprise to find that places with an inordinate amount of crime tend to employ a large police force. Nor is it unreasonable to suspect that jurisdictions increase the size of their police forces when they witness or expect an increase in the level of crime.
- Thus, neither cross-sectional nor time-series analyses can credibly identify a causal effect of police on crime.
- Expenditures on police alone, for example, are over \$65 billion a year.
- The enormous expenditure on policing makes breaking the endogeneity circle more than a mere academic puzzle.

Effect of Police on Crime

- Isolating a causal relationship between increases in the number of police and reductions in the level of crime has large policy consequences.
- In a seminal paper, Steven Levitt showed how the circle could be broken by identifying variations in police presence that were not caused by variations in crime.
- Police presence increased in mayoral and gubernatorial election years but not in off-election years. Since crime is unlikely to be correlated with election timing, this identification strategy can, in principle, break the circle.
- However, this strategy proved to be problematic in practice.
- Variations in police presence brought on by electoral cycles are not large, and variations in other factors impede precise estimation.
- Although Levitt initially did estimate a significant deterrent effect, Justin McCrary later showed that a programming error made Levitt's results appear more precise than justified.

Effect of Police on Crime

- A stronger research design than that used in the past and a new data source allow to better estimate the causal effect of police on crime.
- On March 11, 2002, the Office of Homeland Security introduced the Homeland Security Advisory System (HSAS) to inform the public and other government agencies about the risk of terrorist attacks. During high-alert times, the police increase their presence on the streets of Washington, D.C.
- This variation may be used to break the circle of endogeneity to estimate the effect of police on crime.
- Similar paper to Klick and Tabarrok (2005). Rafael Di Tella and Ernesto Schargrodsky (2004): A terrorist attack on the main Jewish center in Buenos Aires in July 1994 led to increased police presence on blocks with Jewish and Muslim institutions (mosques, synagogues, schools, and so forth). Di Tella and Schargrodsky found that auto theft declined by 75 percent on protected blocks but that little or no changes were observed one or two blocks distant.

Effect of Police on Crime

- Like Di Tella and Schargrodsky (2004), Klick and Tabarrok (2005) take advantage of presumably exogenous shocks to police presence and the fact that these shocks may have different impacts across space and time.
- Data: daily police reports of crime from the Metropolitan Police Department of the District of Columbia (Washington, D.C.).

CRIMES IN WASHINGTON, D.C., BY TYPE:
MARCH 12, 2002–JULY 30, 2003 (506 Days)

Offense Category	Total	Daily Average
Assault with a deadly weapon	5,682	11.2
Arson	129	.3
Burglary	7,071	14.0
Homicide	368	.7
Robbery	5,937	11.7
Sex abuse	530	1.0
Stolen auto	12,149	24.0
Theft	10,230	20.2
Theft from auto	13,726	27.1
Total	55,882	110.4

Effect of Police on Crime

- The HSAS alert system is broken into five color-coded threat conditions: low (green), guarded (blue), elevated (yellow), high (orange), and severe (red).
- Since its inception, the HSAS has never fallen below elevated, but, on four occasions during our time period, it has risen to high, the second highest level. The threat level was high September 10–14, 2002; February 10–27, 2003; March 17–April 16, 2003; and May 20–30, 2003.
- During a high-alert period, the D.C. police department increases the number of patrols, increases the length of shifts in order to put more police on the street, etc.

Effect of Police on Crime

TOTAL DAILY CRIME DECREASES ON HIGH-ALERT DAYS

	(1)	(2)
High Alert	−7.316*	−6.046*
	(2.877)	(2.537)
Log(midday ridership)		17.341**
		(5.309)
R^2	.14	.17

NOTE.—The dependent variable is the daily total number of crimes (aggregated over type of crime and district where the crime was committed) in Washington, D.C., during the period March 12, 2002–July 30, 2003. Both regressions contain day-of-the-week fixed effects. The number of observations is 506. Robust standard errors are in parentheses.

* Significantly different from zero at the 5 percent level.

** Significantly different from zero at the 1 percent level.

Effect of Police on Crime

- Regress daily D.C. crime totals against the terror alert level (1 = high, 0 = elevated) and a day-of-the-week indicator.
- The coefficient on the alert level is statistically significant at the 5 percent level and indicates that on high-alert days, total crimes decrease by an average of seven crimes per day, or approximately 6.6 percent.
- Use dummy variables (not shown) for each day of the week to control for day effects (crime is highest on Fridays).
- Hypothesize that the level of crime decreases on high-alert days in D.C. because of greater police presence on the streets.
- An alternative hypothesis is that tourism is reduced on high-alert days, and as a result, there are fewer potential victims, which leads to fewer crimes.
- To test whether fewer visitors could explain the results, the authors obtained daily data on public transportation (Metro) ridership.

Effect of Police on Crime

- The Metro data suggest a very small decrease in midday ridership on high-alert days.
- Specifically, while days without a high alert averaged 116,000 midday riders, that number decreased to 113,000 riders on high-alert days, a decrease of less than 3 percent.
- To investigate the effect of tourism more systematically, the authors verify that high-alert levels are not being confounded with tourism levels by including logged midday Metro ridership directly in the regression.
- The coefficient on the alert level is slightly smaller, at -6.2 crimes per day. Interestingly, the authors find that increased Metro ridership is correlated with an increase in crime. The increase, however, is very small a 10 percent increase in Metro ridership increases the number of crimes by only 1.7 per day on average.

Effect of Police on Crime

- While suggestive of the effect of police on crime, THE data provide more variation to exploit.
- Washington, D.C., is split into seven police districts.
- The White House, Congress, Smithsonian Institution, and many other prominent government agencies and public areas of Washington, D.C., are located in District 1, the National Mall area.
- One may hypothesize that during a terror alert, most of the increased police attention will be devoted to District 1.
- During periods of high alert, crime in the National Mall area decreases by 2.62 crimes per day.
- Crime also decreases in the other districts, by .571 crimes per day, but this effect is not statistically significant.
- On an average day, there are 17.1 crimes on the National Mall, implying a decline during high-alert days of approximately 15 percent, more than twice as large as that found for the city as a whole.

Effect of Police on Crime

REDUCTION IN CRIME ON HIGH-ALERT DAYS: CONCENTRATION ON THE NATIONAL MALL

	Coefficient (Robust)	Coefficient (HAC)	Coefficient (Clustered by Alert Status and Week)
High Alert × District 1	-2.621** (.044)	-2.621* (1.19)	-2.621* (1.225)
High Alert × Other Districts	-.571 (.455)	-.571 (.366)	-.571 (.364)
Log(midday ridership)	2.477* (.364)	2.477** (.522)	2.477** (.527)
Constant	-11.058** (4.211)	-11.058 (5.87)	-11.058 ⁺ (5.923)

NOTE.—The dependent variable is daily crime totals by district. Standard errors (in parentheses) are clustered by district. All regressions contain day-of-the-week fixed effects and district fixed effects. The number of observations is 3,542. $R^2 = .28$. HAC = heteroskedastic autocorrelation consistent.

⁺ Significantly different from zero at the 10 percent level.

^{*} Significantly different from zero at the 5 percent level.

^{**} Significantly different from zero at the 1 percent level.

Effect of Police on Crime

REDUCTION IN CRIME ON HIGH-ALERT DAYS: CONCENTRATION AMONG STREET CRIMES

	PROPERTY CRIMES			
	VIOLENT CRIMES	Auto Theft and Theft from Auto	Burglary	Theft
High Alert \times District 1 ⁺	-.007 (.373)	-2.383* (.714)	-.288 ⁺ (.171)	.058 (.279)
High Alert \times Other Districts ⁺	-.057 (.116)	-.409 (.267)	-.169 ⁺ (.090)	.065 (.098)
Log(midday ridership)	-2.684 (1.916)	.527 (.363)	.247 ⁺ (.137)	1.128** (.153)
Constant	-.007 (.373)	2.319 (4.064)	-1.284 (1.534)	-9.409** (1.714)
Mean in District 1 during high alert	N.A.	5.56	1.951	N.A.
District 1 high alert/mean \times 100 (%)	N.A.	-42.8	-14.8	N.A.

NOTE.—The dependent variable is daily crime totals by district. All regressions contain day-of-the-week dummies and district fixed effects. All standard errors are heteroskedastic autocorrelation consistent (Newey-West) with clustering by district. N.A. = not applicable.

⁺ Significantly different from zero at the 10 percent level.

^{*} Significantly different from zero at the 5 percent level.

^{**} Significantly different from zero at the 1 percent level.