ØAMET4100 · Spring 2019
Lecture Note 1A

Instructor: Fenella Carpena

January 10, 2019

This lecture note provides an overview of econometrics and reviews random variables and probability theory (Stock & Watson, Chapters 1 and 2). This lecture note is not intended to be a comprehensive review of lecture or the textbook, since there is a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Economic Questions and Data

1.1 Economic Questions We Examine

Many decisions in economics, business, and government require an understanding of relationships among variables in the world around us. These decisions require quantitative answers to quantitative questions. What kinds of questions can we answer using econometrics?

1. Education policy: Does reducing class size improve elementary school education?
2. Racial bias in mortgage lending: Is there racial discrimination in the market for home loans?
3. Tax policy: How much do cigarette taxes reduce smoking?
4. Macroeconomic forecasting: By how much will U.S. GDP grow next year?

The first there of these questions concern causal relationships, while the last question concerns forecasting.

1.2 Causal Effects and Idealized Experiments

Econometric methods are typically used for **causal inference** or **forecasting.**

With **causal inference**, the purpose of the analysis to infer whether one variable (such as eating meat) has a **causal effect** on another variable (such as colon cancer). We typically use **randomized control experiments**—a procedure that randomly assigns units to treatment or control groups—to produce data the reveal causal relationships. The causal effect is the effect on an outcome of a given action or treatment, as measured in an ideal randomized control experiment. In an experiment, the only systematic reason for differences in outcomes between treatment and control groups is the treatment itself. In practice, however, it is very difficult to perform ideal experiments. We will return to a discussion of experiments later in the semester.

Forecasting can be loosely defined as the process of applying a statistical model to *predict new or future observations.* In other words, the goal is to develop a formula to make predictions about one variable (such as next quarter's growth), based on observed values of another (such as this quarter's growth).[1] Forecasting need not involve causal relationships: for instance, one way to make a forecast about whether it is raining is to observe whether pedestrians are using umbrellas, but the act of using an umbrella does not cause it to rain.

[1] As an example, the Santa Cruz Police Department implements "predictive policing," that is deploying officers in places where crimes are likely to occur in the future. In particular, the department fits statistical models to historical crime data to make projections about which areas and windows of time are at highest risk for future crimes. This helps the department deploy resources in a more efficient way, given a tight operating budget. See the NYTimes article about it here: `http://www.nytimes.com/2011/08/16/us/16police.html`.

1.3 Data: Sources and Types

In this course, we will examine data that come from both experimental and non-experimental observations of the world. **Experimental data** come from experiments designed to evaluate a treatment or policy or to investigate a causal effect. **Observational data** are obtained by observing actual behavior outside an experimental setting. Observational data pose major challenges to econometric attempts to estimate causal effects, and the econometric tools we will learn in this course are designed to tackle these challenges. Both experimental or observational data can be one of three types:

- **Cross-sectional data** refer to data on different entities—workers, consumers, firms, governmental units, etc.—for a single time period.
- **Time series data** refer to data for a single entity (e.g., person firm, country) collected at multiple time periods.
- **Panel data**, also called **longitudinal data**, are data for multiple entities in which each entity is observed at two or more periods.

2 Review of Probability

2.1 Random Variables and Probability Distributions

- A **random variable** is a numerical summary of a random outcome. It can be either **discrete** (i.e., can take on only a discrete set of values such as 0, 1, or 2) or **continuous** (i.e., can take on a continuous set of values).[2]
  - ○ *Notation: Upper case letters (e.g., X, Y, Z etc.) are used to denote random variables. We will also use "r.v." as an abbreviation for "random variable."*
  - ○ *Notation: Lower case letters indicate a generic possible value of the random variable.*
- Every random variable X is associated with a **cumulative distribution function (cdf)**. The cdf is concerned with the probability that the random variable takes on a value less than or equal to a particular number. It is defined as $F(x) = P(X \leq x)$.
  - ○ *Notation: The cdf is denoted with upper case F. $P(X \leq x)$ is read as the "probability that X is less than or equal to x," where x is a generic value. For example, we can write $F(4) = P(X \leq 4)$. In some cases, we may also write F with a subscript. For instance, we may write F_X to emphasize that the cdf corresponds to X and not any other random variable.*
- **Important properties of a cdf**: (1) The cdf goes to 0 at negative infinity and goes to 1 at positive infinity; (2) The cdf is a non-decreasing function.
- In addition to the cdf, every random variable is associated with another function, called either the **probability mass function (pmf)** or the **probability density function (pdf)**. The pmf and pdf refer to the discrete and continuous cases, respectively. Both the pmf and the pdf are concerned with "point probabilities" of random variables.
  - ○ *Notation: The pmf or pdf is denoted with lower case f.*
- The **pmf** of a discrete random variable is given by $f(x) = P(X = x)$ for all x. The pmf can be presented in a tabular format, listing all the possible outcomes of the random variable and the probability at which each outcome will occur.
- The **pdf** of a continuous random variable is a function that satisfies $F(x) = \int_{-\infty}^{x} f(t)dt$ for all x. Note that since a continuous random variable can take on an infinite number of values, the probability that it is equal to a particular value is infinitesimally zero.
  - ○ **Important:** If a and b are constants with $a < b$, the area under the pdf between points a and b equals $P(a \leq X \leq b)$. $P(X \leq a)$ is the area under the pdf to the left of the point a, while $P(X \geq b)$ is the area under the pdf to the right of point b.

[2] Mixed random variables exist as well but we are not going to cover them in this course.

- **Important properties of a pdf or pmf**: (1) $f(x) \geq 0$ for all x. This means that the probabilities are all non-negative, since probabilities can never be less than 0; (2) the probabilities sum up to 1 (for discrete) or integrate to 1 (for continuous). This happens because the pmf or pdf must cover all possible states of the world.

- Note that **the pdf/pmf and cdf contain the same information.**

  - For a discrete r.v.. the cdf $F(x)$ is obtained by summing the pdf over all x_j such that $x_j \leq x$.
  - For a continuous r.v., the cdf $F(x)$ is the area under the pdf to the left of the point x.

- **Some useful properties when working with probability:**

  1. For any constant c, $P(X > c) = 1 - P(X \leq c) = 1 - F(c)$.
  2. For any constants $a < b$, $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$.
  3. In the case where X is a continuous r.v., it does not matter whether the inequality in the previous two points are strict or not, i.e., $P(X \geq c) = P(X > c)$ and $P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$. This is because as previously mentioned, for a continuous r.v., the probability that it is equal to a particular value is infinitesimally zero.

2.2 Expected Values, Mean, and Variance

- The **expected value** of a random variable X, denotedx $E(X)$, is a weighted average of all possible outcomes of X, where the weights are the probabilities of that outcome. The expected value of X is also called the **expectation** of X or the **mean** of X. The precise definition of $E(X)$ depends on whether the variable X is discrete or continuous. For most of this lecture note, we will focus on the discrete case.

  - For a discrete r.v. X, $E(X) \equiv \sum_{i=1}^{N} x_i \cdot P(X = x_i)$, where N is the total number of possible values of X
  - *Notation: $E(X)$ may also be denoted as μ. In some cases, we may also write μ with its r.v. as a subscript, e.g. μ_X emphasizes that μ refers the mean of X and not any other r.v.*

- The **variance** of a random variable X, denoted $var(X)$, is the expected value of the squared deviations of X from its mean. It is a measure of how much the distribution of X is tightly centered around its mean. Note that variance is always non-negative.

  - $var(X) \equiv E[(X - \mu)^2] = E(X^2) - \mu^2$.
  - For a discrete r.v. X, $var(x) \equiv E[(X - \mu)^2] = \sum_{i=1}^{k} (x_i - \mu)^2 \cdot P(X = x_i)$.
  - *Notation: We also denote $var(X)$ as σ^2 or σ_X^2.*

- The **standard deviation** of a random variable, denoted $sd(X)$, is the positive square root of the variance. Note that the units of $sd(X)$ are the same as the units of X.

  - $sd(X) \equiv \sqrt{var(X)}$.
  - *Notation: We also denote $sd(X)$ as σ or σ_X.*

2.3 Two Random Variables

- The **joint probability distribution** of two discrete random variables X and Y gives the probability for simultaneous outcomes, $P(X = x, Y = y)$. The probabilities of all possible (x, y) combinations sum to 1. Note that it is also possible to define a joint distribution for continuous random variables (e.g., bivariate normal distribution).

- The **marginal probability distribution** of a random variable Y is another way of referring to its probability distribution. This terminology is used to distinguish the distribution of Y alone (i.e., the marginal distribution of Y) from the joint distribution of Y and another variable X.

  - The marginal distribution of a discrete r.v. Y can be computed as $P(Y = y) = \sum_{i=1}^{l} P(X = x_i, Y = y)$. Here, the value of Y is fixed at y, and we are adding up the probabilities over the l different values that X can take, i.e., $x_1, \ldots, x_l$.

- The distribution of a random variable Y conditional on another random variable X taking on a specific value is called the **conditional distribution** of Y given X.
   - The conditional distribution of Y given $X = x$ is given by $P(Y = y|X = x) = \frac{P(X=x,Y=y)}{P(X=x)}$.
- The **conditional expectation** (or conditional mean) of Y is the mean of the conditional distribution of Y given X.
   - For the discrete case, it is computed as $E(Y|X = x) = \sum_{i=1}^{k} y_i \cdot P(Y = y_i|X = x)$. In this formula, note that the value of X is fixed at x, and we are summing over all k possible values of Y, i.e., $y_1, \ldots, y_k$.
- The **conditional variance** of Y given X is the variance of the conditional distribution of Y given X.
   - For the discrete case, it is computed as $var(Y|X = x) = \sum_{i=1}^{k}[y_i - \mu_{Y|X}]^2 \cdot P(Y = y_i|X = x)$, where $\mu_{Y|X} \equiv E(Y|X = x)$.
- The **law of iterated expectations** states that the expectation of Y is equal to the expectation of the conditional expectation of Y given X.
   - Mathematically, it is stated as $E(Y) = E[E(Y|X)]$. Here, the inner expectation on the right-hand side is computed using the conditional distribution of Y given X, and the outer expectation is computed using the marginal distribution of X.
   - For a discrete random variable X which takes on l values $x_1, \ldots, x_l$, $E(Y) = E[E(Y|X)] = \sum_{i=1}^{l} E(Y|X = x_i)P(X = x_i)$. Hence, another way of saying the law of iterated expectations is that the mean of Y is the weighted average of the conditional expectation of Y given X, where the weights come from the probability that $X = x$ occurs.
   - Example: Suppose you want to compute the mean height of the US population. The law of iterated expectations says that you could compute it by first taking the mean height of males and females separately, and then averaging these two values using the proportion of male and female population as weights. In other words, $E[\text{height}] = E[\text{height}|\text{male}] * P[\text{male}] + E[\text{height}|\text{female}] * P[\text{female}]$.
- Two discrete random variables X and Y are said to be **independent** if and only if $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ for all possible x and y. Intuitively, independence means that knowing the value of X gives us no information about the value of Y, and vice versa. If the random variables are not independent, they are said to be **dependent**.
- The **covariance** between two random variables X and Y, denoted $cov(X, Y)$, is defined as the expected value of the product of $(X - \mu_X)(Y - \mu_Y)$. It measures the amount of *linear* dependence between the two random variables X and Y.
   - $cov(X, Y) \equiv E(X - \mu_X)(Y - \mu_Y) = E(XY) - \mu_X\mu_Y$.
   - *Notation: cov(X,Y) may also be denoted as* σ_{XY}.
- The **correlation** between two random variables X and Y, denoted $corr(X, Y)$, is a scale-free measure of linear dependence between these variables. If $corr(X, Y) = 0$, then X and Y are said to be **uncorrelated**.
   - $corr(X, Y) \equiv \frac{cov(X,Y)}{sd(X)sd(Y)}$
   - If $E(Y|X) = 0$, then $cov(Y, X) = 0$ and $corr(Y, X) = 0$. **Can you show why this is true?**
   - *Notation: corr(X,Y) may also be denoted as* ρ *or* ρ_{XY}.
- **Some important properties of expectation, variance, sd, covariance, and correlation:** In what follows, let X, Y and Z be random variables, and let a, b, c, d be constants. The last two properties listed below are **very important** and will show up again in later parts of the course.
   1. $E(a) = a$
   2. $E(aX + b) = aE(X)$
   3. $E(X + Y) = E(X) + E(Y)$

4. $var(a) = 0$, $sd(a) = 0$, i.e. the variance and sd of any constant is zero.
5. $var(aX + b) = a^2Var(X)$
6. $var(aX + bY) = a^2var(X) + b^2var(Y) + 2ab \cdot cov(X, Y)$
7. $sd(aX + b) = |a|sd(X)$
8. $cov(a, X) = 0$
9. $cov(aX + c, bY + d) = a \cdot b \cdot cov(X, Y)$
10. $cov(X + Y, Z) = cov(X, Z) + cov(Y, Z)$
11. $cov(X, X) = var(X)$
12. $corr(X, Y)$ must be between -1 and 1, i.e. $-1 \leq corr(X, Y) \leq 1$. Values of $corr(X, Y)$ closer to 1 or -1 indicate a stronger linear relationship between X and Y.
13. If X and Y are independent, then:
    - $E(XY) = E(X)E(Y)$
    - $var(X + Y) = var(X) + var(Y)$
14. If X and Y are independent, then $cov(X, Y) = 0$. However, the converse is not true. That is, if $cov(X, Y) = 0$, this does **NOT** imply X and Y are independent.
15. If $cov(X, Y) \neq 0$, then X and Y are not independent.

2.4 The Normal, Chi-Squared, Student t, and F Distributions

2.4.1 The Normal Distribution

The normal distribution is the most widely used distribution in statistics. It is very important for this course because we will rely heavily on it for many topics.

- A normal random variable is a continuous r.v. that can take on any value. We say that X has a **normal distribution** with expected value μ and variance σ^2, written as $X \sim \mathcal{N}(\mu, \sigma^2)$. The normal pdf is symmetric about it mean, μ.

- A special case of the normal distribution occurs when the mean is 0 and the variance is 1. This is called the **standard normal distribution.**

    - If the r.v. Z has $\mathcal{N}(0, 1)$ distribution, then we say Z is a **standard normal random variable.**
    - *Notation: The letter Z is usually used to denote a standard normal r.v.*
    - The pdf of the standard normal r.v. is denoted $\phi(z)$. The pdf of a normal r.v. X is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}}exp(-(x - \mu)^2/2\sigma^2)$ for $-\infty < x < \infty$.
    - The cdf of the normal distribution is denoted $\Phi(z)$. It is obtained as the area under ϕ, to the left of z. Since Φ represents the cdf, $\Phi(z) = P(Z \leq z)$.
    - The values of the probabilities $\Phi(z)$ of a standard normal r.v. are tabulated in the back of your textbook (see *Appendix Table 1*, page 803). It is very important to understand how this table works, since we will be using it over the course of the semester.

- **Important property:** Any normal r.v. X can be transformed into a standard normal r.v.

    - Mathematically, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$. This means that to turn X into a standard normal r.v., we need to subtract its mean μ and divide by the sd σ. This process is also called standardizing the variable.
    - The term $(X - \mu)/\sigma$ is also sometimes called the Z-statistic.
    - Standardizing the variable is important because it allows us to find the probabilities of **any** normal r.v., by first transforming it to a standard normal and then using the *Appendix Table 1* at the back of the textbook.

- **Important formulas to remember when working with a normal random variable Z:**

    1. $P(Z > z) = 1 - P(Z \leq z) = 1 - \Phi(z)$

2. $\Phi(-z) = P(Z < -z) = P(Z > z)$. This holds because of symmetry of the normal normal distribution.
3. $P(|Z| > z) = P(Z > z) + P(Z < -z) = 2P(Z > z) = 2P(Z < -z) = 2\Phi(-z)$ Again, this holds because of symmetry of the normal distribution.
4. For any constants a and b with $a < b$, $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$

- The normal distribution can be generalized to the case to describe the joint distribution of random variables, which is called the **multivariate normal distribution.** In the specific case of only two variables, it is called a **bivariate normal distribution.** The multivariate (and bivariate) normal distributions have four important properties.

1. If X and Y have a bivariate normal distribution with covariance σ_{XY} and if a and b are constants, then $aX + bY \sim N(\mu_X + \mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY})$. More generally, if n random variables have a multivariate normal distribution, any linear combination of these variables is normally distributed.
2. If a set of variables has a multivariate normal distribution, then the marginal distribution of each of the variables is normal.
3. If variables with a multivariate normal distribution have covariances that equal zero, then the variables are independent. This is a special property of the multivariate normal distribution, and it is not generally true that zero covariance implies independence.
4. If X and Y have a bivariate normal distribution, then the conditional expectation of Y given X is linear in X.

2.4.2 Chi-Squared Distribution

- The **chi-squared distribution** is used when testing certain types of hypotheses in statistics and econometrics.

- The chi-squared distribution is the distribution of the sum of m squared independent standard normal random variables. It is denoted as χ^2_m.

2.4.3 Student t Distribution

- The **Student t distribution** with m degrees of freedom is defined to be the distribution of the ratio of a standard r.v., divided by the square root of an independently distributed chi-squared r.v. with m degrees of freedom divided by m.

- Specifically, if Z is a standard normal r.v., W is an r.v. with a chi-squared distributed with m degrees of freedom, and Z and W are independent, then the r.v. $Z = \sqrt{W/m}$ has a Student t distribution with m degrees of freedom.

- The Student t distribution is denoted t_m.

2.4.4 F Distribution

- The F distribution with m and n degrees of freedom (denoted $F_{m,n}$) is the distribution of the ratio of a chi-squared r.v. with degrees of freedom m, divided by m, to an independently distributed chi-squared r.v. with n degrees of freedom n, divided by n.

- Specifically, if W is a chi-squared r.v. with m degrees of freedom, and V is a chi-squared r.v. with n degrees of freedom, then $\frac{W/m}{V/n}$ has an $F_{m,n}$ distribution.

- In econometrics, an important special case of the F distribution is when the denominator degrees of freedom is large enough so that the $F_{m,n}$ distribution can be approximated by the $F_{m,\infty}$ distribution. The $F_{m,\infty}$ distribution is the distribution of a chi-squared r.v. with m degrees of freedom, divided by m. That is, W/m is distributed $F_{m,\infty}$.

- The critical values for the $F_{m,\infty}$ distribution are shown in Appendix Table 4 of the textbook. The 90th, 95th, and 99th percentiles of the $F_{m,n}$ distribution are given in Appendix Table 5 for selected values of m and n. It is very important to understand how these tables work, since we will be using it later in the semester.

2.5 Random Sampling and the Sample Average

- Throughout this course, we will want to answer questions about the **population**, defined as the entire collection of interest or the group of entities (such as people, companies, or school districts) being studied.
    - Example 2.5.1. An online retailer might be interested in understanding how satisfied its customers are with their checkout experience. In this case, the population is everyone in the retailer's customer base.
- However, collecting information about the entire population is usually impractical, infeasible, or very costly. As a result, we typically use a **sample**, a selected subset of the population that we include in our survey. A sample is said to be **representative** if it reflects the mix in the entire population. Samples that distort the population are said to have **bias**. A representative sample can be achieved by **random sampling**, i.e., selecting a subset of the population at random.
    - Example 2.5.2. Continuing from Example 2.5.1, the online retailer might randomly select a sample of 1,000 customers from its database, where each customer has the same likelihood of being chosen for the sample. The retailer can then send a survey to the selected customers, asking them to rate their customer experience from 1 (poor) to 5 (excellent).
- One method of sampling is called **simple random sampling.** This occurs when entities are chosen independently from a population using a method that ensures that each entity is equally likely to be chosen. The process described in Example 2.5.2. is an example of simple random sampling.
- Using the survey data from our sample, we can calculate a **sample average** or **sample mean**, a characteristic observed in the sample.
- We denote the sample mean as $\overline{X}$. It is the average of n individual measurements from a population, where n is our sample size. Mathematically, we state the sample mean formally as $\overline{X} = \frac{X_1+X_2+...+X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Here, note that the value of the random variable X for the i^{th} drawn object is denoted X_i.
    - Example 2.5.3. Continuing from Example 2.5.2, from the survey data we collected from our sample, we can calculate the *average checkout rating of customers* in the sample. Assuming that all surveyed customers responded to the survey, the value of $n = 1000$, and X_1 is the checkout rating of the first customer who was randomly selected for the survey, X_2 is the checkout rating of the second customer who was randomly selected for the survey, etc.
- Because the X_i's (that is, $X_1, \ldots, X_n$) are drawn from the same population, the marginal distribution of X_i is the same for each $i = 1, \ldots, n$. Hence, the X_i's are said to be identically distributed. If, in addition to being identically distributed, the X_i's are independent, they are said to be **independent and identically distributed (i.i.d.).** Let's break down what i.i.d means.
    - *Independent*: Each of the X_i's are independent from each other. For example, X_1 and X_5 are independent, X_2 and X_3 are independent, etc. As a consequence of independence, the covariance of any pair of X_i's is zero. For example, $cov(X_1, X_5) = 0$, $cov(X_2, X_3) = 0$.
    - *Identically Distributed*: All individual measurement X_i have the same distribution. In particular, all X_i's have the same mean μ_X and the same standard deviation σ_X.
    - Simple random sampling produces n random observations $X_1, \ldots, X_n$ that are i.i.d.
- **Important:** The sample average $\overline{X}$ is a random variable and has a sampling distribution, because the value of the sample mean is different every time we take a different sample. In other words, since **there is only one population but many possible samples,** the value of the sample statistic we obtain would differ from sample to sample. This variability is called **sampling variation**.
- Important formulas to know about $\overline{X}$ (make sure you know how to derive these, see equation 2.44 and 2.45 of the textbook):
    - $E(\overline{X}) = \mu_X$
    - $var(\overline{X}) = \frac{\sigma_X^2}{n}$
    - $sd(\overline{X}) = \frac{\sigma_X}{\sqrt{n}}$

2.6 Law of Large Numbers

- The **Law of Large Numbers (LLN)** states that if $X_1, \ldots, X_n$ are i.i.d. with $E(X_i) = \mu_X$ and if large outliers are unlikely, then $\overline{X} \xrightarrow{p} \mu_X$.

- What does $\overline{X} \xrightarrow{p} \mu_X$ mean? This is "convergence in probability." We say that $\overline{X}$ converges in probability to μ_X if for any constant c, the probability that $\overline{X}$ is in the range $(\mu_X - c)$ to $(\mu_X + c)$ becomes close to 1 as n increases.

- Intuitively, the LLN says that as the sample size n increases, with very high probability (approaching 1), the sample average $\overline{X}$ will be arbitrarily close to the expected value μ_X

2.7 Central Limit Theorem

- In general, we do not know what the distribution of the sample mean $\overline{X}$ is.

- However, the **Central Limit Theorem (CLT)** tells us that if (1) $X_1, \ldots, X_n$ are i.i.d. with $E(X_i) = \mu_X$ and $var(X_i) = \sigma_X^2$, (2) $var(X_i) = \sigma_X^2$ is finite, i.e. $var(X_i) = \sigma_X^2 < \infty$, and (3) the sample size n is sufficiently large, then $\overline{X}$ is approximately normally distributed with mean μ_X and variance σ_X^2/n. That is,

$$\overline{X} \sim \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

- The CLT is a powerful result that underpins most of modern applied econometrics. **Why do we care about the CLT?** We care because even if we do not know what the distribution of $\overline{X}$ is, the CLT tells us that in large samples, it is normally distributed, and we can use this fact to conduct hypothesis tests of $\overline{X}$, as we will see in the next chapter.