DATA 5690 Class Notes

Computational Methods for FinTech

Dr. Tyler J. Brough

2024-01-11

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

Introduction

This is a book created from markdown and executable code.

See Knuth for additional discussion of literate programming.

Part I

Microfoundations

1 Summary

In summary, this book has no content whatsoever.

Introduction

These notes on ***probability*** are based on Chapter 2 of Martinez and Martinez (2016).

We start with the concept of an ***experiment***, which is defined as a *process or action whose outcome cannot be predicted with certainty and would likely change when the experiment is repeated.*

- This is the process by which the data are generated in the real-world.
- **NB:** It does not necessarily imply that the agent has complete or even partial experimental control.

Next we define the ***sample space*** denoted by S.

Part II

Arbitrage

2 Option Pricing Theory

... when judged by its ability to explain the empirical data, option pricing theory is the most successful theory not only in finance, but in all of economics. – Stephen Ross

References

Martinez, Wendy L., and Angel R. Martinez. 2016. *Computational Statistics Handbook with Matlab*. Third. CRC Press.

A Elementary Probability Review

This is a review of elementary probability that will be useful for our study of asset pricing. It is based on coverage in Greene, as well as Wooldridge.

Observational data sets econometrics apart from statistics. We will view an economic variable as an **outcome** from a **random process** not under the control of the researcher.

The descriptive term for this underlying mechanism is the **data–generating process**, or **DGP**.

We view the outcome variable X as a random variable because until it is observed we are not certain about its value.

An **experiment** is a procedure that can (at least in theory) be infinitely repeated and has a well–defined set of outcomes.

An example: flip a coin 10 time and count the number of heads. Each time the experiment is repeated the outcome will be an integer between 0 and 10.

A **random variable** is a variable that takes on numerical values and has an outcome that is determined by an experiment.

An example:

- An airline wants to decide how many reservations to book for a flight with 100 seats.
- If fewer than 100 people want reservations they should book them all.
- If more than 100 people want reservations a safe bet may be to only book 100. But not everyone will show up, resulting in lost revenue.
- If they book too many they will have to compensate passengers for having to bump them.

By convention, random variables are denoted by uppercase variables, such as X, Y, and Z.

The corresponding outcomes are denoted by lowercase letters x, y, z.

In the coin flipping example X denotes the number of heads in 10 flips. We don't know ahead of time what value X will take, but we know it will be in the set $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. A particular outcome may be $x = 7$.

Random variables are defined to take on numerical values. So even in the coin flipping example, where the outcomes are "heads" and "tails", we code the outcomes as follows:

- $X = 1$ for heads (success)
- $X = 0$ for tails (failure)

A random variable that only takes on values 0 and 1 is a **Bernoulli** (or **binary**) **random variable**.

A discrete random variable is one that takes on only a finite (or countably infinite) number of values. A binary variable is the simplest case of a discrete random variable. The only quantity that we need to completely describe its behavior is the probability that $X = 1$.

In the coin flipping example (if the coin is "fair") then

- $P(X = 1) = \frac{1}{2}$
- $P(X = 0) = \frac{1}{2}$

Consider again the airline's problem of booking seats on a flight. We can analyze this with several binary variables. For a randomly selected passenger define a binary variable as $X = 1$ if she shows up for the flight, and $X = 0$ otherwise. There is no reason to believe in this case that $P(X = 1) = \frac{1}{2}$, so we will define a *parameter* θ so that:

- $P(X = 1) = \theta$
- $P(X = 0) = 1 - \theta$

For example, if $\theta = 0.75$, then there is a 75% chance of the passenger showing up and 25% chance of not showing up. In a real–life business situation the actual value of θ is crucial in determining the airline's strategy.

Methods for *estimating* θ, given historical data on airline reservations is the subject of *mathematical statistics*.

Generally, a discrete random variable is completely described by listing the set of possible outcomes and the associated probability that it takes on each value.

If X has k possible values $\{x_1, x_2, \ldots, x_k\}$ then the probabilities $p_1, p_2, \ldots, p_k$ are defined by:

- $p_j = P(X = x_j)$, for $j = 1, 2, \ldots, k$
- $0 \leq p_j \leq 1$
- $\sum_{j=1}^{k} p_j = 1$

The **probability function** or (**PDF**) of X summarizes the information concerning the possible outcomes of X and the corresponding probabilities:

$$f(x_j) = p_j, \quad j = 1, 2, \ldots, k$$

For any real number x, $f(x)$ is the probability that the random variable X takes on the particular value x.

An example: Suppose that X is the number of free throws made by Larry Bird out of two attempts. X can take on the three values $\{0, 1, 2\}$. Assume the PDF of X is given by

- $f(0) = 0.20$
- $f(1) = 0.44$
- $f(2) = 0.36$

We can calculate the probability that Larry Bird will make at least one free throw:

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) \\ &= 0.44 + 0.36 \\ &= 0.80 \end{aligned}$$

We can graph this discrete PDF as follows:

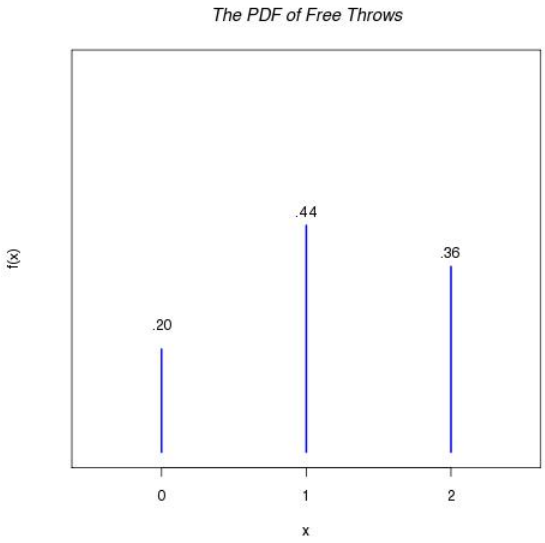

Figure A.1: The Probability Density Function of Larry Bird Free Throws

When dealing with more than two random variables we subscript the PDF's as follows:

- f_x is the PDF of X
- f_y is the PDF of Y

A variable X is a **continuous random variable** if it takes on any real value with *zero* probability. A continuous random variable X can take on so many possible values that they are not countable, so logical consistency requires that each one has probability zero.

Examples:

- Prices
- Wages
- Interest rates
- Height
- Weight
- Waiting time

The **cumulative distribution function** (**CDF**) of the random variable X is:

$$F(X) = P(X \leq x)$$

For discrete random variables it is obtained by summing the PDF over all values x_j such that $x_j \leq x$.

For a continuous random variable, $F(X)$ is the area under the PDF, $f(x)$ to the left of x.

Because it is a probability, $0 \leq F(X) \leq 1$.

If $x_1 < x_2$ then $P(X \leq x_1) \leq P(X \leq x_2)$, that is $F(x_1) \leq F(x_2)$.

Two important properties of CDFs that are useful for computing probabilities are the following:

- For and number c, $P(X > c) = 1 - F(c)$
- For any numbers a and b, $P(a \leq X \leq b) = F(b) - F(a)$

For continuous random variables the inequalities in probability statements are not strict:

$$P(X \geq c) = P(> c)$$

$$\begin{aligned} P(a < X < b) &= P(a \leq X \leq b) \\ &= P(a \leq X < b) \\ &= P(a < X \leq b) \end{aligned}$$

Let X and Y be discrete random variables. Then for (X, Y) a **joint distribution** which is fully described by the **joint probability density function** of (X, Y):

$$f_{XY}(x, y) = P(X = x, Y = y)$$

X and Y are said to be independent if, and only if:

$$f_{XY}(x, y) = f_X(x) f_Y(y) \quad \text{for every } x \text{ and } y$$

where f_X is the PDF of the random variable X, and f_Y is the PDF of random variable Y.

f_X and f_Y are referred to as the **marginal probability density functions**.

The discrete case is the easiest to grok. If X and Y are discrete and independent then

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

Note: If X and Y are independent then finding the joint PDF only requires knowledge of $P(X = x)$ and $P(Y = y)$

Example: Consider a basketball player shooting two free throws. Let X be the Bernoulli random variable equal to 1 if he makes the first free throw, and 0 otherwise. Let Y be the Bernoulli random variable equal to 1 if he makes the second free throw. Suppose that he is an 80% free throw shooter, so that $P(X = 1) = P(Y = 1) = 0.80$. What is the probability of making both free throws?

If X and Y are independent: $P(X = 1, Y = 1) = P(X = 1) P(Y = 1) = (0.8)(0.8) = 0.64$. Thus, a 64% chance of making both.

Independence is often reasonable in more complicated situations. In the airline example, suppose that n is the number of reservations booked. For each $i = 1, 2, \ldots, n$ let Y_i denote the Bernoulli random variable indicating whether or not customer i shows up for the flight.

Let θ again denote the probability of success (showing up for the reservation). Each $Y_i \sim Bern(\theta)$.

The variable of primary interest is the total number of customers showing up out of the n reservations: call this X.

$$X = Y_1 + Y_2 + ... + Y_n$$

Assume that $P(Y_i = 1) = \theta$ for every Y_i, and further that they Y_i are independent. Then X has a **binomial distribution**, which we write in shorthand as: $X \sim Bin(n, \theta)$. The binomial PDF is the following:

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \quad \text{for } x = 0, 1, 2, ..., n$$

Note: $\binom{n}{x} = \frac{n!}{x!(n-x)!}$, and is read as "n choose x".

Example: If the flight has 100 seats and $n = 120$ and $\theta = 0.85$ then:

$$P(X > 100) = P(X = 101) + P(X = 102) + ... + P(X = 120)$$

In econometrics we are usually interested in how one variable Y is related to one or more other variables. For now, consider only one such variable X. What we can know about how X affects Y is contained in the **conditional distribution** of Y given X. This information is summarized in the **conditional probability distribution function**:

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

In the discrete case: $f_{Y|X}(y|x) = P(Y = y|X = x)$, which we read as the probability that $Y = y$ given that $X = x$.

If X and Y are independent, then the knowledge of X tells us nothing about Y:

$$f_{Y|X}(y|x) = f_Y(y) \quad \text{and}$$
$$f_{X|Y}(x|y) = f_X(x)$$

Example: Free throw shooting again. Assume the conditional PDF is given by the following:

- $f_{Y|X}(1|1) = 0.85$, and $f_{Y|X}(0|1) = 0.15$.
- $f_{Y|X}(1|0) = 0.70$, and $f_{Y|X}(0|0) = 0.30$.

These are not independent. The probability of making the second free throw depends on whether or not the first free throw was made. We can calculate $P(X = 1, Y = 1)$ if we know $P(X = 1)$. Assume the probability of making the first free throw is $P(X = 1) = 0.80$. Then:

$$\begin{aligned} P(X = 1, Y = 1) &= P(Y = 1|X = 1) \times P(X = 1) \\ &= (0.85) \times (0.80) \\ &= 0.68 \end{aligned}$$

The **expected value** is a measure of central tendency. It is one of the most important probabilistic concepts in econometrics. If X is a random variable the **expected value** (or expectation) of X, denoted $E(X)$ and sometimes μ, is a weighted average of all possible values of X. The weights are determined by the PDF.

Consider the case of a discrete random variable. Let $f(x)$ denote the PDF of X. The expected value of X is the weighted average:

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + ... + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j)$$

Example: Suppose X takes on the values -1, 0, and 2 with probabilities $\frac{1}{8}$, $\frac{1}{2}$, $\frac{3}{8}$. Then

$$E(X) = (-1)(\frac{1}{8}) + (0)(\frac{1}{2}) + (2)(\frac{3}{8}) = \frac{5}{8}$$

Note: $E(X)$ can take on values that are not even possible outcomes of X.

If X is a continuous random variable then

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

This is still interpreted as a weighted average.

Given a random variable X and a function $g(\cdot)$, we can create a new random variable $g(X)$. For example, if X is a random variable, then so is X^2 or $log(X)$ (for $x > 0$).

The expected value of $g(X)$ is given by

$$E[g(X)] = \sum_{j=1}^{k} g(x_j) f_X(x_j)$$

or

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

Example: For the random variable above let $g(X) = X^2$. Then

$$E(X^2) = (-1)^2(\frac{1}{8}) + (0)^2(\frac{1}{2}) + (2)^2(\frac{3}{8}) = \frac{13}{8}$$

Note: $E[g(X)] \neq g[E(X)]$.

Properties of Expected Values:

- **Property E1:** For any constant c, $E(c) = c$.
- **Property E2:** For any constants a and b, $E(aX + b) = aE(X) + b$.
- **Property E3:** If $a_1, a_2, ..., a_n$ are constants and $X_1, X_2, ..., X_n$ are random variables then:
  - $E(a_1X_1 + a_2X_x + ... + a_nX_n) = a_1E(X_1) + a_2E(X_2) + ... + a_nE(X_n)$
  - Or $E(\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} a_iE(X_i)$
  - A special case is when each $a_i = 1$ so that $E(\sum_{i=1}^{n} E(X_i)) = \sum_{i=1}^{n} E(X_i)$, or in other words the expected value of a sum, is the sum of the expected values.

Example: Expected revenue at a pizzeria. X_1, X_2, and X_3 are the number of small, medium, and large pizzas sold during the day. Suppose $E(X_1) = 25$, $E(X_2) = 57$, and $E(X_3) = 40$. Prices are \$5.50 for a small, \$7.60 for a medium, and \$9.15 for a large. Then expected revenue is the following

$$\begin{aligned} E(5.50X_1 + 7.60X_2 + 9.15X_3) &= 5.50E(X_1) + 7.60E(X_2) + 9.15E(X_3) \\ &= 5.50(25) + 7.60(57) + 9.15(40) \\ &= 936.70 \end{aligned}$$

The outcome on any given day will differ from this, but this is the expected revenue.

If $X \sim Bin(n, \theta)$ then $E(X) = n\theta$. The expected number of successes in n Bernoulli trials is $n\theta$. We can see this by writing

$$X = Y_1 + Y_2 + ... + Y_n \quad \text{where each} Y_i \sim Bern(\theta)$$

Then

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} E(Y_i) \\ &= \sum_{i=1}^{n} \theta \\ &= n\theta \end{aligned}$$

Example: Consider the airline problem with $n = 120$ and $\theta = 0.85$. Then $E(X) = n\theta = 120(0.85) = 102$, which is too many.

The **median** is another measure of central tendency. If X is continuous then the median is the value m such that one–half of the area under the PDF is to the left of m, and one–half is to the right of m.

If X is discrete and takes on an odd number of finite values, the median is obtained by ordering the possible outcomes of X and selecting the middle value.

Example: For the sample $\{-4, 0, 2, 8, 10, 13, 17\}$ the median is 8.

If X takes on an even number of values, then the median is the average of the two middle values.

Example: For the sample $\{-5, 3, 9, 17\}$ the median is $\frac{3+9}{2} = 6$.

For a random variable let $E(X) = \mu$. There are various ways to measure how far X is from its expected value. One of the simplest is the squared distance:

$$(X - \mu)^2$$

This eliminates the sign, which corresponds with our intuitive notion of a distance measure. It treats values above and below μ symmetrically.

The **variance** is defined as follows:

$$Var(X) = E[(X - \mu)^2]$$

The variance is sometimes denoted by σ_X^2 or just σ^2 when the random variable is understood to be X.

Note:

$$\begin{aligned}\sigma^2 &= E(X^2 - 2X\mu + \mu^2) \\ &= E(X^2) - 2\mu^2 + \mu^2 \\ &= E(X^2) - \mu^2\end{aligned}$$

Example: If $X \sim Bern(\theta)$ we know that $E(X) = \theta$. Since $X^2 = X$ it follows that $E(X^2) = \theta$. Then $Var(X) = E(X^2) - \mu^2 = \theta - \theta^2 = \theta(1-\theta)$.

Properties of variance:

- **Property VAR1:** $Var(X) = 0$ if, and only if for every c such that $P(X = c) = 1$, in which case $E(X) = c$.
- **Property VAR2:** For constants a and b $Var(aX + b) = a^2 Var(X)$.

The **standard deviation** is related to the variance as follows: $sd(X) = \sqrt{Var(x)}$. The standard deviation is often denoted σ_x or just σ.

Properties of the standard deviation:

- **Property SD1:** For a constant c, $sd(c) = 0$.
- **Property SD2:** For constants a and b $sd(aX + b) = |a|sd(X)$.

Given a random variable X, we can define a new random variable Z by

$$Z = \frac{X - \mu}{\sigma}$$

or $Z = aX + b$ where $a = \frac{1}{\sigma}$ and $b = \frac{-\mu}{\sigma}$. Then $E(Z) = aE(X) + b = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$.

The variance is $Var(Z) = a^2 Var(X) = \frac{\sigma^2}{\sigma^2} = 1$. Thus the new random variable has $\mu = 0$ and $\sigma^2 = 1$. This is known as **standardizing** a random variable.

Example: Suppose $E(X) = 2$ and $Var(X) = 9$ then $Z = \frac{X-2}{3}$.

While the joint distribution completely describes the relationship between two random variables it is often useful to have a summary measure of how, on average, two random variables vary with one another.

The **covariance** is defined as follows:

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

The covariance is often denoted by σ_{XY}. If $\sigma XY > 0$ then on average when X is above its mean Y is also above its mean. If $\sigma_{XY} < 0$ then on average when X is above its mean Y is below its mean, and vice versa.

Note:

$$\begin{aligned} Cov(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[(X-\mu_X)Y] \\ &= E(XY) - \mu_X\mu_Y \end{aligned}$$

Properties of covariance:

- **Property COV1:** If X and Y are independent then $Cov(X,Y) = 0$. Note: the converse is not true. Zero $Cov(X,Y)$ does not imply independence.
- **Property COV2:** For any constants a_1, b_2, a_2, and b_2 $Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X,Y)$.
- **Property COV3:** $|Cov(X,Y)| \leq sd(X)sd(Y)$.

Note: property COV2 suggests that $Cov(X,Y)$ depends upon how the random variables are measured, not only on how strongly they are related. In other words, scale matters for $Cov(X,Y)$.

The **correlation coefficient** is defined as

$$Corr(X,Y) = \frac{Cov(X,Y)}{sd(X)sd(Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

The correlation coefficient is sometimes denoted by ρ_{XY}.

Properties of correlation:

- **Property CORR1:** $-1 \leq Corr(X,Y) \leq 1$.
- **Property CORR2:** For constants a_1, b_1, a_2, and b_2 with $a_1a_2 > 0$ $Corr(a_1X+b_1, a_2Y+b_2) = Corr(X,Y)$. If $a_1a_2 < 0$ then $Corr(a_1X + b_1, a_2Y + b_2) = -Corr(X,Y)$.

With covariance and correlation defined we state further properties of the variance:

- **Property VAR3:** For constants a and b, $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X,Y)$.
- **Property VAR4:** If $\{X_1, X_2, \ldots, X_n\}$ are pairwise uncorrelated and $\{a_i : i = 1, \ldots, n\}$ are constants then $Var(\sum_{i=1}^{n} a_iX_i) = \sum_{i=1}^{n} a_i^2Var(X_i)$.

The **conditional mean** is defined as follows:

$$E(Y|x) = \sum_{j=1}^{m} y_j f_{Y|X}(y_j|x)$$

Example: Let (X, Y) represent the population of all working individuals, where X is years of education and Y is hourly wages. Then $E(Y|X = 12)$ is the average hourly wage for all the people in the population with 12 years of education (roughly high school education). $E(Y|X = 16)$ is the average hourly wage for all people with 16 years of education.

A typical situation in econometrics will look like the following:

$$E(WAGE|EDUC) = 1.05 + 0.45EDUC$$

If this linear relationship holds then for 8 years of education the expected hourly wage is $1.05 + 0.45(8) = 4.65$ of \$4.65 per hour.

Properties of conditional expectations:

- **Property CE1:** $E[c(X)|X] = c(X)$ for any function $c(X)$. In other words, functions act as constants. For example, $E[X^2|X] = X^2$. If we know X we also know X^2.
- **Property CE2:** For funtions $a(X)$ and $b(X)$, $E[a(X)Y+b(X)|X] = a(X)E(Y|X)+b(X)$. For example, consider the random variable $XY + 2X^2$. $E(XY + 2X^2|X) = XE(Y|X) + 2X^2$.
- **Property CE3:** If X and Y are independent then $E(Y|X) = E(Y)$.
- **Property CE4:** $E[E(Y|X)] = E(Y)$. This is known as the Law of Iterated Expectations.
- **Property CE5:** $E(Y|X) = E[E(Y|X, Z)|X]$.
- **Property CE6:** If $E(Y|X) = E(Y)$ then $Cov(X, Y) = 0$ and $Corr(X, Y) = 0$.

The **conditional variance** is defined as follows:

$$Var(Y|X = x) = E(Y^2|X) - [E(Y|X)]^2$$

Properties of conditional variance:

- **Property CV1:** If X and Y are independent then $Var(Y|X) = Var(Y)$.

The **normal probability density function** is defined as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \frac{-(X-\mu)^2}{2\sigma^2}, \quad \text{for } -\infty < x < \infty$$

where $E(X) = \mu$ and $Var(X) = \sigma^2$. When is a random variable is normally distributed we write $X \sim N(\mu, \sigma^2)$.

A special case is the **standard normal distribution**, which is defined as follows:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp \frac{-z^2}{2}, \quad \text{for } -\infty < z < \infty$$

The standard normal cumulative distribution function is denoted by $\Phi(z) = P(Z \leq z)$. Using some basic facts from probability we arrive at the following helpful formulas:

$$\begin{aligned} P(Z > z) &= 1 - \Phi(z) \\ P(Z < -z) &= P(Z > z) \\ P(a \leq Z \leq b) &= \Phi(b) - \Phi(a) \end{aligned}$$

Properties of the normal distribution:

- **Property NORMAL1:** If $X \sim N(\mu, \sigma^2)$ then $\frac{(X-\mu)}{\sigma} \sim N(0, 1)$.
- **Property NORMAL2:** If $X \sim N(\mu, \sigma^2)$ then $aX + b \sim N(a\mu + b, a^2\sigma^2)$.
- **Property NORMAL3:** If X and Y are jointly normally distributed, then they are independent if, and only if $Cov(X, Y) = 0$.
- **Property NORMAL4:** Any linear combination of independent, identically distributed normal random variables has a normal distribution.

Example: Let X_i for $i = 1, 2, \text{and}, 3$, be independent random variables distributed as $N(\mu, \sigma^2)$. Define $W = X_1 + 2X_2 - 3X_3$. Then W is normally distributed. We can solve for the mean and variance as follows:

$$\begin{aligned} E(W) &= E(X_1) + 2E(X_2) - 3E(X_3) = \mu + 2\mu - 3\mu = 0 \\ Var(W) &= Var(X_1) + 4Var(X_2) + 9Var(X_3) = 16\sigma^2 \end{aligned}$$

The **chi–square distribution** is obtained directly from independent, standard normal random variables. Let Z_i, $i = 1, 2, \ldots, n$, be independent random variables, each distributed as standard normal. Define a new random variable as the sum of the squares of the individual Z_i:

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

The new random variable X has a **chi–square distribution** with n **degrees of freedom**. This is often written as $X \sim \chi_n^2$.

The t **distribution** is a workhorse in classical statistics and econometrics. A t distribution is obtained from a standard normal and a chi–square random variable. Let Z have a standard normal distribution and let X have a chi-square distribution with n degrees of freedom. Also assume that Z and X are independent. Then the following random variable

$$T = \frac{Z}{\sqrt{Z/n}}$$

has a t distribution with n degrees of freedom. This is denoted by $T \sim t_n$. The t distribution gets its degrees of freedom from the chi–square random variable.

Another important distribution for statistics and econometrics is the F **distribution**. To define an F random variable, let $X_1 \sim \chi_{k_1}^2$ and $X_2 \sim \chi_{k_2}^2$ and assume that X_1 and X_2 are independent. Then, the random variable

$$F = \frac{X_1/k_1}{X_2/k_2}$$

has an F distribution with (k_1, k_2) degrees of freedom. We denote this as $F \sim F_{k_1,k_2}$. The order of the degrees of freedom is important. k_1 is the *numerator degrees of freedom* and k_2 is the *denominator degrees of freedom*.

B Mathematical Statistics Review

B.1 Introduction

This is a review of fundamental mathematical statistics that will be essential for learning econometrics. The coverage is based on Wooldridge.

Statistical inference is the process of learning something about a population given a sample from that population. Using the tools of statistics we will seek to *infer* something about the population, given only a sample.

A **population** is a well defined group of subjects, such as individuals, firms, cities, etc.

By learning, we mainly mean two things:

- Estimation
- Hypothesis Testing

An example of a population is all working adults in the US. Labor economists are interested in learning about the return to education, measured by the average increase in earnings given another year of education. It is impractical or impossible to gather data on the entire population, but she can obtain data on a subset of the population. Using the data collected a labor economist may report that her best estimate of the return to another year of education is 7.5%. This is an example of a **point estimate**. Or she may report a range, such as "the return to education is between 5.6% and 9.4%." This is an example of an **interval estimate**.

An urban economist might want to know whether neighborhood crime watch programs are associated with lower crime rates. After comparing crime rates of neighborhoods with and without such programs in a sample from the population, he can draw one of two conclusions: neighborhood watch programs do affect crime, or they do not. This is an example of **hypothesis testing**.

The first step in statistical inference is to identify the population of interest. Once a population has been identified, a model for the population relationship of interest may be specified. Models involve probability distributions or features of probability distributions, and these depend on unknown parameters. **Parameters** are constants that determine the directions and strengths of relationships among variables. In the labor economics example, the parameter of interest is the return to education in the population.

B.2 Sampling

Let Y be a random variable representing a population with PDF $f(y;\theta)$, which depends on a single parameter θ. The PDF is assumed to be known, except for θ. Different values of θ imply different population distributions, and therefore we are interested in θ. If we can obtain samples from the population we can learn something about θ.

Random sampling:

If $Y_1, Y_2, ... Y_n$ are independent random variables with a common PDF $f(y;\theta)$ then $\{Y_1, Y_2, \dots, Y_n\}$ is said to be a *random sample* from $f(y,\theta)$ (a random sample represented by $f(y;\theta)$).

When $\{Y_1, Y_2, \dots, Y_n\}$ is a random sample from $f(y,\theta)$, we also say that the Y_i are **independent and identically distributed** (or iid) random variables from $f(y;\theta)$.

If family income is obtained for $n = 100$ families in the US, the incomes we observe will differ for each sample of 100 that we choose. Once a sample is obtained we have a set of number $\{y_1, y_2, \dots, y_3\}$, which constitute the data that we work with.

Random samples from Bernoulli distributions are often used to illustrate statistical concepts. If $Y_1, Y_2, \dots, Y_n$ are iid Bernoulli(θ), such that $P(Y_i = 1) = \theta$ and $P(Y_i = 0) = 1 - \theta$ then $\{Y_1, Y_2, \dots, Y_n\}$ constitute a random sample from a Bernoull(θ) distribution.

Consider the airline example: Each Y_i denotes whether or not passenger i shows up. θ is the probability that a randomly drawn individual from the population shows up.

For many applications, random samples can be assumed to be drawn from a normal distribution. If $\{Y_1, Y_2, \dots, Y_n\}$ is a random sample from the Normal(μ, σ^2) population, the population is characterized by two parameters, the mean mu and the variance σ^2.

Finite sample properties are properties that hold for a sample of any size, no matter how small or large (sometimes called "small sample properties" to distinguish from "asymptotic properties").

B.3 Estimation and Estimators

Given a random sample drawn from a population distribution that depends on an unknown parameter θ. An **estimator** of θ is a rule that assigns each possible outcome of the sample a value of θ. The rule is specified before any sampling is carried out (regardless of the data collected).

Let $\{Y_1, Y_2, \dots, Y_n\}$ be a random sample from a population with mean μ. A natural estimator of μ is the average of the random sample:

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

$\bar{Y}$ is called the **sample average**; unlike earlier when we defined the average as a descriptive statistics, $\bar{Y}$ is now viewed as an estimator. Given any outcome of the random variables $Y_1, Y_2, \ldots, Y_n$, we use the same rule to estimate μ: we average them. For actual outcomes $\{y_1, y_2, \ldots, y_n\}$, the estimate is just the average in the sample.

$$\bar{y} = \frac{(y_1 + y_2 + \cdots + y_n)}{n}$$

More generally an estimator W of a parameter θ can be expressed as:

$$W = h(Y_1, Y_2, \ldots, Y_n)$$

for some known function h of the random variables $Y_1, Y_2, \ldots, Y_n$. W is a random variable because it depends on the random sample: as we obtain different random samples from the population, the value of W can change.

When a particular set of numbers $\{y_1, y_2, \ldots, y_n\}$ is plugged into h, we obtain an *estimate* of θ, denoted $w = h(y_1, y_2, \ldots, y_n)$.

So we have that:

- W is a point estimator
- w is a point estimate

to evaluate estimation procedures we study various properties of the PDF of W. The distribution of an estimator is called its **sampling distribution**. In mathematical statistics, we study the sampling distributions of estimators.

Unbiasedness: an estimator, W of θ, is an unbiased estimator if

$$E(W) = \theta$$

for all possible values of θ.

Remarks:

- If an estimator is unbiased, then its PDF has an expected value equal to the parameter it is estimating. However, in any given sample $E(W)$ may not equal θ.
- Rather, if we could indefinitely draw samples from the population, getting an estimate each time, and then average these estimates over all random samples we would obtain θ.

- This is just a thought experiment, because in reality we have only one sample to work with. But this "what if" property is desirable for estimators.

If W is a **biased estimator** of θ, its bias is defined as

$$Bias(W) = E(W) - \theta$$

Example: $\bar{Y}$ is an unbiased estimator of the population mean, μ

$$\begin{aligned} E(\bar{Y}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E(Y_i)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu \end{aligned}$$

Example: s^2 is an unbiased estimator of σ^2.

Let $\{Y_1, Y_1, \ldots, Y_n\}$ denote a random sample from the population with

- $E(Y) = \mu$
- $Var(Y)\sigma^2$

then

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

This is usually called the **sample variance**.

Note: the division by $n-1$ accounts for the fact that μ is estimated by $\bar{Y}$ and not known. If μ were known $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$, would be an unbiased estimator.

Unbiasedness has some weaknesses:

- Some very reasonable, even very good, estimates are not unbiased.
- Some unbiased estimates are quite poor.

For example:

$$W = Y_1 (\text{i.e. discard all other observations})$$

It is an unbiased estimator $E(Y_1) = \mu$.

Example: If $n = 100$, we have one hundred observation of the random variable Y, but we discard all but the first to estimate $E(Y)$.

The weaknesses of unbiasedness show that we need additional criteria to evaluate estimators. Unbiasedness ensures that the sampling distribution of an estimator has a mean value equal to the parameter it is estimating.

We also want to know how spread out it is. The variance of an estimator is called its **sampling variance** because it is the variance associated with the sampling distribution.

Example:

$$\begin{aligned} Var(\bar{Y}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} Var(Y_i)\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma^2\right) = \frac{1}{n^2} n\sigma^2 = \frac{1}{n}\sigma^2 \end{aligned}$$