Harvard Econometrics Math Camp Notes[1]

Summer 2021[2]

These notes provide a brief review of several topics that will be useful to think about before entering the first-year econometrics sequence at Harvard.

DISCLAIMERS:

1. If some of the material in these notes is unfamiliar to you, *do not worry*! You will have plenty of time to brush up this year.
2. These notes contain more content than we will have time to cover during math camp. This is intentional. We hope that these notes can be a useful reference material for you throughout the year.

Chapters

[1] These notes are heavily based on textbooks and notes from other sources. These include Gary Chamberlain's lecture notes, Paul Sally's *Tools of the Trade*, and the graduate Scientific Computing course (AM205) at Harvard.

[2] Ashesh Rambachan created the first draft of this document in 2018. Frank Pinter provided revisions in 2019. Michael Droste provided revisions in 2020 and 2021, and he takes responsibility for all errors.

Chapter 1. Review of Linear Algebra

These notes provide a brief review of linear algebra. Although we focus primarily on applications of linear algebra to econometrics, you will find the tools developed in the study of linear algebra to be useful throughout your first year.

Contents

IF PROBABILITY AND STATISTICS are the foundation of econometrics, linear algebra is something closer to a toolbox. Vectors and matrices are an unavoidable part of the work of econometrics, and vector and matrix operations figure prominently in our formulas and proofs. As soon as we begin talking about vector-valued random variables, these operations grab a foothold in our work. Vectors and matrices also come up in situations like the following:

- We store data in the computer using data vectors and data matrices.
- Statistical software (for instance, Mathematica, R, Matlab, NumPy, and Julia) are optimized for matrix calculations, so speaking the language of matrices makes our code run more quickly.[1]
- Expressing systems of equations in terms of vectors and matrices is much more parsimonious than writing out each equation separately.

Linear algebra also lets us think at a higher level of abstraction, and simplify our calculations and proofs. To raise the level of abstraction yet higher, we use tools that conceptually unify the operations we perform on finite data (sample mean, variance, covariance, ordinary least squares) with their population analogs (expectation, variance, covariance, population regression).

In these notes, we will present results more generally than you probably saw in linear algebra, but more concretely than any course in functional analysis would. We do this partly to allow the jump to random variables to come naturally, but also to give a flavor of how deep these results are.

This chapter is not intended to replace a good course in linear algebra. In particular, they are targeted towards concepts that occur often in the first-year econometrics sequence. Many standard topics are not covered, or are only covered briefly[2]. In addition, these notes omit proofs, which are an essential part of learning the material. We recommend Hoffman and Kunze and/or Fraleigh for an overview of linear algebra, Luenberger for inner product spaces and the projection theorem, and Golub and Van Loan for matrix decompositions.

Preliminaries

Linear algebra is all about vectors and matrices. Vector spaces are built on some very primitive mathematical structures. It will be necessary to (very briefly) develop some machinery from abstract algebra, groups and fields, in order to think properly about the properties of vector spaces. We will do that now - bear with us!

Definition 1.1. *A **commutative group** (or **abelian group**), denoted $(G, *)$, is a set G and a binary operation $* : G \times G \to G$ satisfying:*

1. *Closure: for all $a, b \in G$, we have $a * b \in G$.*
2. *Associativity: for all $a, b, c \in G$, we have $(a * b) * c = a * (b * c)$.*
3. *Commutativity: for all $a, b \in G$, we have $a * b = b * a$.*
4. *Identity element: there exists $e \in G$ such that, for all $a \in G$, $a * e = a$.*
5. *Inverses: for each $a \in G$, there exists $a' \in G$ such that $a * a' = e$.*

Remark 1.1. *The study of groups is incredibly elegant and interesting. Certain sets, when combined with certain mathematical operations (e.g. addition, multiplication), will naturally have a "group structure".*

[1] Computational linear algebra is a large field, and all empirical economists owe it an enormous debt. For a taste, consult Golub and Van Loan, or take Scientific Computing in the applied math department (AM 205).

[2] For instance, eigenvalues and eigenvectors are essential to the study of dynamical systems and dynamic models, but will be covered very briefly for the purposes of this chapter.

Definition 1.2. *A **field**, denoted $(F, +, \cdot)$, is a set F and two operations $+ : F \times F \to F$ and $\cdot : F \times F \to F$ satisfying:*[3]

1. *$(F, +)$ forms a commutative group. We denote the additive identity element by $0 \in F$ and the additive inverse of $a \in F$ by $-a$.*
2. *Closure under $\cdot$: for all $a, b \in F$, we have $a \cdot b \in F$.*
3. *Associativity of $\cdot$: for all $a, b, c \in F$, we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.*
4. *Commutativity of $\cdot$: for all $a, b \in F$, we have $a \cdot b = b \cdot a$.*
5. *Multiplicative identity element: there exists $1 \in F$ such that, for all $a \in F$, $a \cdot 1 = a$.*
6. *Multiplicative inverses for nonzero elements: for each $a \in F \setminus \{0\}$, there exists $a^{-1} \in F$ such that $a \cdot a^{-1} = 1$.*
7. *Distributivity: for all $a, b, c \in F$, $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.*

The rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers[4] $\mathbb{C}$ are all fields. This note generally focuses on the real numbers here, though the results will generalize to the complex numbers.

Vector spaces

Definition 1.3. *A **vector space** over a field F, denoted $(V, +, \cdot)$, is a set V and two operations $+ : V \times V \to V$ and $\cdot : F \times V \to V$ satisfying:*

1. *$(V, +)$ forms a commutative group. We denote the additive identity element by $0 \in V$ and the additive inverse of $v \in V$ by $-v$.*
2. *Closure under $\cdot$: for all $a \in F$ and $v \in V$, we have $a \cdot v \in V$.*
3. *For all $a \in F$ and $v_1, v_2 \in V$, we have $a \cdot (v_1 + v_2) = (a \cdot v_1) + (a \cdot v_2)$.*
4. *For all $a, b \in F$ and $v \in V$, we have $(a + b) \cdot v = (a \cdot v) + (b \cdot v)$.*
5. *For all $a, b \in F$ and $v \in V$, we have $(ab) \cdot v = a \cdot (b \cdot v)$.*
6. *For all $v \in V$, we have $1 \cdot v = v$.*

Elements of a vector space are called **vectors**, elements of the field are called **scalars**, and the operation $\cdot$ is called **scalar multiplication**.[5] The most familiar example is $\mathbb{R}^n$, considered as a vector space over $\mathbb{R}$; we will usually treat our data vectors as elements.[6]

Exercise 1.1. *Let X and Y be random variables. Show that $\{\alpha X + \beta Y \mid \alpha, \beta \in \mathbb{R}\}$ is a vector space over $\mathbb{R}$ with the usual addition and scalar multiplication operations for random variables. What is the additive identity element?*

Definition 1.4. *Let $(V, +, \cdot)$ be a vector space over a field F, and let $W \subseteq V$. We say W is a **subspace** of V if W is closed under addition and scalar multiplication.*

Dimension and basis

Let V be a vector space over a field F, let $v_1, \ldots, v_n \in V$, and let $\alpha_1, \ldots, \alpha_n \in F$. Then

$$\alpha_1 v_1 + \cdots + \alpha_n v_n$$

[3] The $\cdot$ in multiplication is usually dropped for expediency - something you likely already do all the time.

[4] Numbers of the form $a + b\sqrt{-1}$, where a (the 'real part') and b (the 'imaginary part') are real numbers.

[5] The $\cdot$ in scalar multiplication is usually dropped. Society has agreed that it is best we all understand ab to mean $a \cdot b$, for just about any a, b, or $\cdot$.

[6] Some authors, such as Luenberger, distinguish between the space of points in $\mathbb{R}^n$ and the space of n-dimensional Euclidean vectors, which they label $\mathbb{E}^n$.

is a **linear combination** of the vectors $v_1, \ldots, v_n$. This is an enormously useful object. In particular, once we have a notion of a basis, we can uniquely represent every vector as a linear combination of basis vectors. These linear combinations are more concrete and easier to work with than the abstract vectors they represent, especially if the basis has nice properties.

Definition 1.5. *Let V be a vector space over a field F, and let $v_1, \ldots, v_n$ be nonzero vectors. The set $\{v_1, \ldots, v_n\}$ is **linearly independent** if, for any $\alpha_1, \ldots, \alpha_n \in F$,*

$$\alpha_1 v_1 + \cdots + \alpha_n v_n = 0 \implies \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0.$$

Equivalently, the set is linearly independent if no element can be written as a linear combination of the other elements.

Definition 1.6. *Let V be a vector space over a field F, and let $v_1, \ldots, v_n \in V$. The set $\{v_1, \ldots, v_n\}$ **spans** V if, for any $v \in V$, v can be written as a linear combination of $\{v_1, \ldots, v_n\}$.[7] That is, there exist $\alpha_1, \ldots, \alpha_n \in F$ such that $v = \alpha_1 v_1 + \cdots + \alpha_n v_n$.*

We also sometimes refer to the **span** of $\{v_1, \ldots, v_n\}$, the set of vectors $v \in V$ that can be written as a linear combination of $\{v_1, \ldots, v_n\}$. The span of a set of vectors forms a subspace of V.

Definition 1.7. *Let V be a vector space over a field F and let $B \subseteq V$. We say B is a **basis** for V if B is linearly independent and spans V.*

Definition 1.8. *Let V be a vector space over a field F. We say V is **finite-dimensional** if there exists a finite $S \subseteq V$ that spans V.*

Theorem 1.1. *Let V be a nonzero finite-dimensional vector space over a field F. Then V has a finite basis, and every basis of V has the same number of elements.*

Definition 1.9. *Let V be a nonzero finite-dimensional vector space over a field F. Let B be a set of n vectors that forms a basis of V. Then the **dimension** of V, written $\dim V$, is equal to n.*

If $V = \{0\}$ then define $\dim V = 0$.

Every linearly independent set in V has at most $\dim V$ elements, and every set that spans V has at least $\dim V$ elements. Think of a basis as the (non-unique) smallest set that spans V.[8]

Example 1.1. *Consider $V = \mathbb{R}^n$ as a vector space over $\mathbb{R}$ with the usual addition and scalar multiplication operations. The **standard basis** $\{e_1, \ldots, e_n\}$, where e_j is has 1 in the jth coordinate and 0 everywhere else, is a basis for V, so $\dim V = n$.*

Let V be an n-dimensional vector space over a field F, and let $B = \{v_1, \ldots, v_n\}$ be a basis of V.[9] We can write $v \in V$ uniquely as a linear combination of basis vectors,

$$v = \sum_{i=1}^{n} \alpha_i v_i,$$

where $\alpha_1, \ldots, \alpha_n$ are the **coefficients** of v relative to the basis B. We conventionally write the coefficients in a **column vector** or $n \times 1$ matrix, and call this the **representation of v with respect to the basis** B:[10]

[7] This definition requires n to be finite. In general, a set S spans V if every $v \in V$ can be written as a linear combination of finitely many elements of S.

[8] In a finite-dimensional vector space, any linearly independent set that is not a basis can be completed to form a basis.

Similarly, any set that spans V and is not a basis can drop elements until it becomes linearly independent.

[9] Throughout this note, we will write $B = \{v_1, \ldots, v_n\}$ as if B were a set. Since the order of the basis vectors matters, B is actually a sequence (or an **ordered basis**).

[10] Hoffman and Kunze call this the **coordinate matrix of v with respect to the ordered basis** B and denote it $[v]_B$.

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}.$$

In the case of $\mathbb{R}^n$, representation with respect to the standard basis coincides with the standard coordinate representation.[11]

Example 1.2. *Consider $\mathbb{R}^2$ as a vector space over $\mathbb{R}$. You can show that $B = \{(1,0),(1,1)\}$ is a basis of $\mathbb{R}^2$. Let $(a,b) \in \mathbb{R}^2$; its representation with respect to B is*

$$\begin{pmatrix} a-b \\ b \end{pmatrix}.$$

Vector spaces that are not finite-dimensional are called **infinite-dimensional**. The existence of a basis for every infinite-dimensional vector space is equivalent to the axiom of choice.[12]

Linear Transformations and Matrices

Every matrix represents a function that maps one vector space to another. When thinking of the properties of matrices, I find it helpful look for a geometric intuition in terms of that function.[13]

Definition 1.10. *Let V and W be vector spaces over a field F. A **linear transformation** is a function $T : V \to W$ satisfying:*

1. *For all $v_1, v_2 \in V$, $T(v_1 + v_2) = T(v_1) + T(v_2)$.*
2. *For all $v \in V$ and $\alpha \in F$, $T(\alpha v) = \alpha T(v)$.*

The definition above states that linear transformations are (1) closed under addition and (2) closed under scalar multiplication. To build intuition for this result, here is a simple exercise:

Example 1.3. *Consider a function $T : \mathbb{R} \to \mathbb{R}, T(x) = a + bx$, where $a, b \in \mathbb{R}$. We would like to check whether T is a linear transformation. First, let's check whether T is closed under addition. Note that $T(v_1 + v_2) = a + b(v_1 + v_2) \neq a + bv_1 + a + bv_2 = T(v_1) + T(v_2)$, except when $a = 0$. Next, let's check whether T is closed under scalar multiplication. Note that $T(\alpha v_1) = a + \alpha b v_2 \neq \alpha T(v_1) = \alpha a + \alpha b v_2$, except when $a = 0$. So if $a = 0$, then T is closed under addition and scalar multiplication and is therefore a linear transformation. If $a \neq 0$, then T is neither closed under addition nor scalar multiplication and is not a linear transformation.*

Matrix representation

A linear transformation between two vector spaces can be represented as a matrix, by writing down what the transformation does to each basis vector. Taking a step back, this is a very powerful result: linear transformations can be thought of as being represented by matrices. Matrices are nice to work with, and so that means linear transformations are nice to work with.

[11] We can think of column vector representation as building an alternative coordinate system, with the basis vectors as building blocks.

[12] The set of polynomials with real coefficients is an infinite-dimensional vector space over $\mathbb{R}$. Can you think of a basis for it?

[13] "You should be aware of the fact that an $m \times n$ matrix A with entries a_{ij} is more than just a static array of mn numbers. It is dynamic. It can act." —Charles Pugh

Let V and W be vector spaces over a field F, let $\{v_1, \dots, v_n\}$ be a basis for V, and let $\{w_1, \dots, w_m\}$ be a basis for W. For $k = 1, \dots, n$, consider the coefficients of $T(v_k)$ relative to the basis $\{w_1, \dots, w_m\}$:

$$T(v_k) = \sum_{j=1}^{m} a_{jk} w_j.$$

Then, since any $v \in V$ can be written as $v = \sum_{k=1}^{n} b_k v_k$, we can write $T(v)$ as a linear combination of $\{T(v_1), \dots, T(v_n)\}$:

$$T(v) = \sum_{k=1}^{n} b_k T(v_k) = \sum_{k=1}^{n} b_k \sum_{j=1}^{m} a_{jk} w_j = \sum_{j=1}^{m} \left(\sum_{k=1}^{n} b_k a_{jk} \right) w_j.$$

This gives us the coefficients of $T(v)$ relative to $\{w_1, \dots, w_m\}$. The results that follow just repeat this result using matrix multiplication.

Definition 1.11. *Let V and W be vector spaces over a field F, let $\{v_1, \dots, v_n\}$ be a basis for V, and let $\{w_1, \dots, w_m\}$ be a basis for W. Let $T : V \to W$ be a linear transformation. The **matrix representation** of T with respect to $\{v_1, \dots, v_n\}$ and $\{w_1, \dots, w_m\}$[14] is the $m \times n$ matrix with scalar entries*

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

where the kth column, $(a_{1k}, \dots, a_{mk})$, contains the coefficients of $T(v_k)$ relative to $\{w_1, \dots, w_m\}$.[15]

The matrix representation is unique given a basis of V and a basis of W. We will say the ijth entry of a matrix is a_{ij}, the entry in row i and column j. We sometimes write the set of $m \times n$ matrices with entries in F as $F^{m \times n}$. An $n \times n$ matrix is called **square**.

Definition 1.12. *Let $A \in F^{m \times p}$ with entries a_{ij} and let $B \in F^{p \times n}$ with entries b_{jk}. **Matrix multiplication** is defined as*[16]

$$A \cdot B = C \text{ where } C \in F^{m \times n} \text{ with entries } c_{ik} = \sum_{j=1}^{p} a_{ij} b_{jk}.$$

Matrix multiplication represents function composition.

Theorem 1.2. *Let V and W be vector spaces over a field F, let $\{v_1, \dots, v_n\}$ be a basis for V, and let $\{w_1, \dots, w_m\}$ be a basis for W. Let $T : V \to W$ be a linear transformation, and let A be its matrix representation with respect to $\{v_1, \dots, v_n\}$ and $\{w_1, \dots, w_m\}$. For all $v \in V$, the representation of $T(v)$ with respect to $\{w_1, \dots, w_m\}$ is $A \cdot v$.*

Properties of linear transformations

Though some results use the language of matrices, remember that linear transformations are always working in the background.[17]

Definition 1.13. *Let V be a vector space over a field F. The **identity function** $I : V \to V$ is the linear transformation defined by $I(v) = v$.*

[14] When $V = W$, we will typically use the same basis for the domain and the range.

[15] Seeing the column vector of v as an $n \times 1$ matrix makes more sense if you consider F as a vector space over itself, with basis $\{1\}$, and consider a linear transformation $T : F \to V$ with $T(1) = v$.

[16] The $\cdot$ in matrix multiplication is usually dropped.

[17] To see these results in full generality using the language of linear transformations, see Hoffman and Kunze.

Let $\dim V = n$. The matrix representation of I, with respect to any basis, is the $n \times n$ **identity matrix**, denoted I_n, with $a_{ij} = 1(i = j)$.

Definition 1.14. *Let V be a vector space over a field F, and let $T : V \to V$ be a linear transformation. The **inverse** of T, if it exists, is the function $T^{-1} : V \to V$ such that*

$$T^{-1}(T(v)) = I(v) = v.$$

*If T^{-1} exists, then T is said to be **invertible**.*

Proposition 1.1. *If T^{-1} exists, then T^{-1} is a linear transformation.*

Let $\dim V = n$ and let A be the matrix representation of T with respect to some basis. The **inverse matrix** of A, denoted A^{-1}, is the matrix representation of T^{-1} with respect to the same basis, and

$$A^{-1}A = I_n.$$

If A^{-1} exists, then A is said to be invertible.[18]

We will now give some of the (many) conditions that are equivalent to the invertibility of T.

Definition 1.15. *Let V and W be vector spaces over a field F and let $T : V \to W$ be a linear transformation. The **null space** (or **kernel**) of T is the set of vectors that T maps to the zero vector, $\{v \in V \mid T(v) = 0\}$. If V is finite-dimensional, the **nullity** of T is the dimension of the null space of T.*

Definition 1.16. *If V is finite-dimensional, the **rank** of T is the dimension of the subspace $T(V) \subseteq W$.*

Theorem 1.3. *Rank-nullity theorem*

Let V and W be vector spaces over a field F and let $T : V \to W$ be a linear transformation. Suppose that V is finite-dimensional. Then

$$\operatorname{rank}(T) + \operatorname{nullity}(T) = \dim V.$$

The notion of rank for linear transformations corresponds to the familiar notion of rank for matrices.

Definition 1.17. *Let $A \in F^{m \times n}$, where F is a field. Write the jth column as $v_j = (a_{1j}, \dots, a_{mj}) \in F^m$. The **column space** of A is the subspace of F^m spanned by the columns, $\{v_1, \dots, v_n\}$, and the **column rank** of A is the dimension of its column space.*

*Likewise, write the ith row of A as $w_i = (a_{i1}, \dots, a_{in}) \in F^n$. The **row space** of A is the subspace of F^n spanned by the rows, $\{w_1, \dots, w_m\}$, and the **row rank** of A is the dimension of its row space.*

Proposition 1.2. *Let V and W be vector spaces over a field F and let $T : V \to W$ be a linear transformation. Let A be a matrix representation of T; then $A \in F^{m \times n}$. The rank of T is equal to the row rank of A and to the column rank of A.*

It follows that the row rank and the column rank are equal. We call this the rank of A.

Theorem 1.4. *Let V be a vector space over a field F with $\dim V = n$ and let $T : V \to V$ be a linear transformation. Let A be the matrix representation of T with respect to a basis. The following are equivalent:*

[18] Calculate the matrix inverse using `solve` in R, `inv` in Matlab, and `np.linalg.inv` in NumPy. To solve a system of linear equations $Ax = b$, instead use `solve(A, b)` in R, `A\b` in Matlab, and `np.linalg.solve(A, b)` in NumPy.

1. *T is invertible.*
2. *T is **full rank**; that is,* $\text{rank}(T) = n$.
3. *T is **nonsingular**; that is,* $\text{nullity}(T) = 0$.
4. *A is an invertible matrix.*
5. *A is full rank; that is,* $\text{rank}(A) = n$.
6. *A is nonsingular; that is, $Av = 0$ implies $v = 0$.*

As a result, 'nonsingular' is sometimes used interchangeably with 'invertible'.

Transpose

The familiar notion of a matrix transpose is described below.[19]

Definition 1.18. *Let $A \in F^{m \times n}$ with entries a_{ij}. The **transpose** of A, denoted A' or A^T, is the $n \times m$ matrix whose ijth entry is a_{ji}.*

A square matrix A for which $A' = A$ is called **symmetric**. You can show that, if A is $m \times p$ and B is $p \times n$, then $(AB)' = B'A'$.

Diagonal and triangular matrices

Definition 1.19. *Let $A \in F^{m \times n}$ with entries a_{ij}. A is **diagonal** if $a_{ij} = 0$ when $i \neq j$.*

A square $n \times n$ diagonal matrix takes the form

$$\begin{pmatrix} c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_n \end{pmatrix}.$$

It represents the linear transformation $T : V \to V$ that scales each basis vector v_i by the scalar c_i. It is clearly symmetric, and if all the $c_i \neq 0$, then the inverse is the diagonal matrix with diagonal entries $c_1^{-1}, c_2^{-1}, \ldots, c_n^{-1}$. In this case, the system of linear equations $Ax = b$ is easy to solve: $x_i = c_i^{-1} b_i$.

The computational simplicity remains if we add entries above or below the diagonal, but not both.[20]

Definition 1.20. *Let $A \in F^{n \times n}$ with entries a_{ij}.*

*A is **lower triangular** if all entries above the diagonal are zero, that is, $a_{ij} = 0$ when $i < j$.*

*A is **upper triangular** if all entries below the diagonal are zero, that is, $a_{ij} = 0$ when $i > j$.*

A diagonal matrix is both lower and upper triangular. The product of upper triangular matrices is upper triangular, and the product of lower triangular matrices is lower triangular.

[19] See Sections 3.5–3.7 of Hoffman and Kunze for the definition of the transpose of a linear transformation.

[20] Lower triangular systems are easy to solve by forward substitution, and upper triangular systems are easy to solve by back substitution. See Golub and Van Loan.

Trace, determinant, and eigenvalues

Trace and determinant occur all the time in matrix computations and proofs, but will not recur for the rest of this note. Eigenvalues, a fascinating and rich topic in linear algebra, will only be discussed briefly.

Definition 1.21. *Let $A \in F^{n \times n}$ and label its entries by a_{ij}. The **trace** of A, denoted $\operatorname{trace}(A)$, is the sum of the diagonal entries, $\sum_{i=1}^{n} a_{ii}$.*

Definition 1.22. *If $B \in F^{n \times n}$ is invertible and $A = B^{-1}CB$, then A and C are **similar**.*

If two matrices are similar, they have the same trace. In addition, if A is $m \times n$ and B is $n \times m$, then $\operatorname{trace}(AB) = \operatorname{trace}(BA)$.

To get intuition for the determinant, consider $A \in \mathbb{R}^{n \times n}$. Label its columns $v_1, \dots, v_n$ and picture the convex hull of its columns, $\{\alpha_1 v_1 + \cdots + \alpha_n v_n \mid \alpha_i \in [0,1]\}$. This is a figure in $\mathbb{R}^n$ (called a parallelepiped), and the determinant of A is its oriented volume.

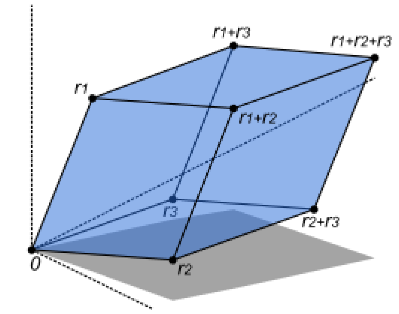

Figure 1: Parallelepiped in $\mathbb{R}^3$. (Claudio Rocchini, Wikimedia Commons)

Definition 1.23. *Let $A \in F^{n \times n}$ with entries a_{ij}. The **determinant** of A, denoted $\det A$ or $|A|$, is defined as follows:*[21]

- *If $n = 1$, then $\det A = a_{11}$.*
- *For $n > 1$, let A_{ij} be the $(n-1) \times (n-1)$ matrix generated by deleting the ith row and jth column of A. Then, letting $i = 1$,*

$$\det A = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det A_{ij}.$$

Proposition 1.3. *The determinant is invariant to the choice of i, as long as $1 \leq i \leq n$.*

Familiar facts about determinants can be guessed from the picture.

- $\det I_n = 1$. In general, the determinant of a diagonal matrix is the product of the diagonal entries.
- $\det A \neq 0$ if and only if A is invertible.[22]
- If A and B are square, then $\det AB = \det A \cdot \det B$. It follows that if A is invertible, then $\det A^{-1} = 1/\det A$.
- $\det A = \det A'$.

Next we discuss eigenvectors and eigenvalues, which measure the ways that a linear transformation shrinks, grows, flips, or deforms the shapes it acts on.

Definition 1.24. *Let $A \in F^{n \times n}$. An **eigenvalue** of A is a scalar $\lambda \in F$ for which there exists some $v \in F^n \setminus \{0\}$, called an **eigenvector**, such that*

$$Av = \lambda v.$$

Any nonzero scalar multiple of v is also an eigenvector corresponding to λ.

Theorem 1.5. *The eigenvalues of A are the roots of the **characteristic polynomial** of A,*

$$p(x) = \det(A - xI_n).$$

[21] See Hoffman and Kunze, Chapter 5, for other equivalent characterizations.

[22] The dimension of the parallelepiped is the column rank of A; if A is not full rank, then the parallelepiped is a zero-measure set in $\mathbb{R}^n$.

Suppose $F = \mathbb{C}$. Then, because p is an n-order polynomial over $\mathbb{C}$, it follows from the fundamental theorem of algebra that p has n complex roots, counted with multiplicity.[23] We can label the eigenvalues (repeating with multiplicity) by $(\lambda_1, \ldots, \lambda_n)$.

The following result is surprisingly useful.

Theorem 1.6. *Let $A \in \mathbb{C}^{n \times n}$. Then*

$$\operatorname{trace}(A) = \sum_{i=1}^{n} \lambda_i \quad \textit{and} \quad \det A = \prod_{i=1}^{n} \lambda_i .$$

Exercise 1.2. *Prove that if two matrices are similar, then they have the same characteristic polynomial, and therefore the same eigenvalues.*

The eigendecomposition anticipates the matrix decompositions we will consider later. It is sometimes helpful in computations and proofs, and gives insight into the workings of a linear transformation.[24]

Definition 1.25. *Let $A \in F^{n \times n}$. A is* ***diagonalizable*** *if there exists a basis of F^n whose elements are all eigenvectors of A.*

For A to be diagonalizable, it is sufficient that (1) A has n distinct eigenvalues, or (2) A is symmetric with real entries.[25]

Theorem 1.7. *Eigendecomposition (or spectral decomposition)*

Let $A \in \mathbb{C}^{n \times n}$ be diagonalizable. Write its eigenvalues, with multiplicity, as $(\lambda_1, \ldots, \lambda_n)$. Then $A = BCB^{-1}$, where C is the diagonal matrix with diagonal elements $\lambda_1, \ldots, \lambda_n$, and the columns of B are eigenvectors of A corresponding to those eigenvalues.

Inner Products

From now on, we will restrict ourselves to vector spaces over $\mathbb{R}$.

Definition 1.26. *An* ***inner product space*** *(or* ***pre-Hilbert space****) over $\mathbb{R}$ is a vector space V over $\mathbb{R}$ together with a function $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$, known as an* ***inner product*** *(or* ***positive definite symmetric bilinear form****), satisfying:*[26]

1. *Symmetry: for all $v, w \in V$, $\langle v, w \rangle = \langle w, v \rangle$.*
2. *Bilinearity (I): for all $v_1, v_2, w \in V$, $\langle v_1 + v_2, w \rangle = \langle v_1, w \rangle + \langle v_2, w \rangle$.*
3. *Bilinearity (II): for all $v, w \in V$ and $\alpha \in \mathbb{R}$, $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$.*
4. *Positive definiteness: for all $v \in V$, $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$.*

Definition 1.27. *A* ***normed linear vector space*** *over a field F is a vector space V over F together with a function $\| \cdot \| : V \to F$, known as a* ***norm****, satisfying:*

1. *For all $v \in V$, $\|v\| \geq 0$, with $\|v\| = 0$ if and only if $v = 0$.*
2. *Triangle inequality: for all $v, w \in V$, $\|v + w\| \leq \|v\| + \|w\|$.*
3. *For all $v \in V$ and $\alpha \in F$, $\|\alpha v\| = |\alpha| \|v\|$.*

Theorem 1.8. *Let V be an inner product space over $\mathbb{R}$. Then the function $\| \cdot \| : V \to \mathbb{R}$ defined by $\|v\| = \sqrt{\langle v, v \rangle}$ is a norm.*

[23] A root λ has multiplicity k if we can factor $p(x) = (x - \lambda)^k s(x)$, where $s(x)$ is a polynomial and $s(\lambda) \neq 0$. The multiplicity of the root is called the **algebraic multiplicity** of the corresponding eigenvalue.

[24] Calculate the eigendecomposition using `eigen` in R, `eig` in Matlab, and `np.linalg.eig` in NumPy.

[25] If $A \in \mathbb{C}^{n \times n}$ is symmetric and every entry has imaginary part zero, then all its eigenvalues have imaginary part zero.

[26] The notion of an inner product space over $\mathbb{C}$ is also well-defined, if we replace symmetry with conjugate symmetry. All the results below hold for inner product spaces over $\mathbb{R}$ or $\mathbb{C}$, so I will usually drop "over $\mathbb{R}$".

We make this distinction because some normed linear vector spaces have norms that do not come from an inner product. The proof of the triangle inequality uses the following result:

Theorem 1.9. *Cauchy-Schwarz inequality*

Let V be an inner product space. For all $v, w \in V$, we must have that $|\langle v, w\rangle| \leq \|v\|\|w\|$.

Furthermore, equality holds if and only if one of the vectors is a scalar multiple of the other.

Dot product and positive definiteness

In the geometry of $\mathbb{R}^n$, the familiar dot product is an inner product. It provides much of the geometric intuition we will lean on for inner product spaces.

Example 1.4. *$\mathbb{R}^n$ with the dot product is an inner product space over $\mathbb{R}$. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$. The* ***dot product*** *of x and y, denoted $x \cdot y$ or $x'y$,[27] is $x_1y_1 + \cdots + x_ny_n$.*

The norm induced by the dot product has a nice interpretation as the length or Euclidean distance. The dot product itself also has a geometric interpretation. Let $x, y \in \mathbb{R}^n$, and let θ be the angle between x and y ($0° \leq \theta \leq 180°$). Then

$$\cos\theta = \frac{x \cdot y}{\|x\|\|y\|}.$$

Definition 1.28. *Let $A \in \mathbb{R}^{n\times n}$.*

A is ***positive semidefinite*** *if, for all $x \in \mathbb{R}^n$, $x \cdot Ax \geq 0$.*

A is ***positive definite*** *if, for all nonzero $x \in \mathbb{R}^n$, $x \cdot Ax > 0$.*

That is, the angle between x and Ax is (weakly) less than $90°$. Objects of the form $x'Ax$ are called **quadratic forms**.

Proposition 1.4. *If A is symmetric and positive definite, then its eigenvalues are positive. If A is symmetric and positive semidefinite, then its eigenvalues are nonnegative.*

Generalizations of dot product

The dot product generalizes conveniently to inner products in infinite-dimensional vector spaces over $\mathbb{R}$. These are useful when we move to sequences, functions, and random variables.

Example 1.5. *Consider infinite sequences of numbers.[28] Define l_2 as the space of real-valued infinite sequences (x_n) where $\sum_{i=1}^{\infty} |x_i|^2 < \infty$, with the inner product as an infinite dot product,*

$$\langle (x_n), (y_n)\rangle = \sum_{i=1}^{\infty} x_i y_i.$$

Example 1.6. *Now consider well-behaved functions.[29] For any $a < b$ define $L_2[a, b]$ as the space of functions $x : [a, b] \to \mathbb{R}$ for which $|x(t)|^2$ is Lebesgue integrable.[30] The inner product is a Lebesgue integral,*

$$\langle x, y\rangle = \int_a^b x(t)y(t)\, dt.$$

[27] We snuck in matrix multiplication here, by representing x and y as column vectors with respect to the standard basis.

[28] This generalizes to discrete random variables with (countably) infinite support.

[29] This generalizes to well-behaved continuous random variables.

[30] Because the Lebesgue integral is not sensitive to the behavior of x on zero-measure sets, we need to consider two functions x and y the same if they are the same almost everywhere (they only differ on a set of measure zero).

Orthogonality

The notion of orthogonality is useful in two ways. First, it is closely connected to linear projection. Second, an orthogonal set (and particularly an orthonormal set) makes for a convenient basis.

Definition 1.29. *Let V be an inner product space and let $v, w \in V$. We say v and w are **orthogonal** (or **perpendicular**) if $\langle v, w \rangle = 0$.*

A vector v is orthogonal to a set $W \subseteq V$ if $\langle v, w \rangle = 0$ for all $w \in W$.

Definition 1.30. *Let V be an inner product space. We say $\{v_1, \ldots, v_n\} \subseteq V$ is an **orthonormal set** if it is pairwise orthogonal (that is, $\langle v_i, v_j \rangle = 0$ whenever $i \neq j$) and $\langle v_i, v_i \rangle = 1$ for all i.*

If $n = \dim V$, then $\{v_1, \ldots, v_n\}$ forms an **orthonormal basis** of V. The standard basis is an example.[31]

Definition 1.31. *Let $Q \in \mathbb{R}^{n \times n}$. Q is an **orthogonal matrix** if its columns[32] form an orthonormal basis of $\mathbb{R}^n$ with the dot product.*

If Q is an orthogonal matrix, then

$$Q'Q = QQ' = I_n.$$

That is, $Q' = Q^{-1}$. Orthogonal matrices represent functions that perform rotation or reflection. In particular, they preserve length:

$$\|Qv\|^2 = (Qv)'Qv = v'Q'Qv = v'v = \|v\|^2.$$

Linear Projections

In econometrics we are often faced with problems of the form

$$\widehat{v} = \arg\min_{w \in W} \|v - w\|$$

asking us to find the best approximation to a vector.[33] The projection theorem gives us simple conditions that guarantee existence and uniqueness of solutions. Furthermore, it gives us a simple way to compute them as the solution to a system of equations.[34]

The best approximation problem asks us to find the projection of a vector on a subspace. From geometry in $\mathbb{R}^n$, we have an intuition that the line connecting the projection with the original vector should be perpendicular to the subspace. Start with the projection of one vector onto another.

Definition 1.32. *Let V be an inner product space over $\mathbb{R}$ and let $v, w \in V$. The **projection** of v onto w is defined by*

$$\text{proj}_w(v) = \frac{\langle v, w \rangle}{\|w\|^2} w.$$

This is the closest point to v in the subspace spanned by w. Furthermore, $v - \text{proj}_w(v)$ is orthogonal to w.

[31] As is any rotation or reflection thereof.

[32] We could equivalently define it using the rows.

[33] For example, ordinary least squares.

[34] No calculus or second order conditions required.

Gram–Schmidt orthogonalization

This is the insight that allows us to generate a basis of orthonormal vectors for any finite-dimensional inner product space. The **Gram–Schmidt orthogonalization procedure** takes any finite linearly independent set of vectors, labeled $\{v_1, \dots, v_n\}$, and generates an orthonormal set that spans the same subspace. It works by iteratively projecting v_k on the vectors that came before, keeping only the orthogonal residual. For $k = 1, \dots, n$:

$$\begin{aligned} z_1 &= v_1, & e_1 &= \frac{z_1}{\|z_1\|} \\ z_2 &= v_2 - \langle v_2, e_1 \rangle e_1, & e_2 &= \frac{z_2}{\|z_2\|} \\ &\vdots \\ z_k &= v_k - \sum_{i=1}^{k-1} \langle v_k, e_j \rangle e_j, & e_k &= \frac{z_k}{\|z_k\|}. \end{aligned}$$

Notice the similarity to residual regression.[35]

Proposition 1.5. *Let $\{v_1, \dots, v_n\}$ be a set of linearly independent vectors in an inner product space V. For each $k = 1, \dots, n$, the subspace spanned by $\{v_1, \dots, v_k\}$ is also spanned by the orthogonal set $\{z_1, \dots, z_k\}$ and by the orthonormal set $\{e_1, \dots, e_k\}$.*

Uniqueness of the projection

Here we generalize from projections onto one-dimensional subspaces to multidimensional subspace W. If the projection is possible, the fact that $v - \text{proj}_w(v)$ is orthogonal to w generalizes nicely.

Theorem 1.10. *Let V be an inner product space, let W be a subspace of V, and let $v \in V$. If there exists $\widehat{v} \in W$ such that $\|v - \widehat{v}\| \leq \|v - w\|$ for all $w \in W$, then $\widehat{v}$ is unique.*

A necessary and sufficient condition for $\widehat{v}$ to be the unique minimizing vector is that $v - \widehat{v}$ is orthogonal to W.

The orthogonality condition, $\langle v - \widehat{v}, w \rangle = 0$ for all $w \in W$, is often much easier to solve than the original minimization problem.

Hilbert spaces and the projection theorem

We need further assumptions to guarantee that $\widehat{v}$ exists. The tools we need should be familiar from real analysis.

Definition 1.33. *Let X be a space equipped with a norm $\| \cdot \|$. A sequence (x_n) of elements in X is **Cauchy** if, for each $\epsilon > 0$, there exists an N such that $\|x_n - x_m\| \leq \epsilon$ for all $n, m > N$.*

Definition 1.34. *We say X is **complete** if every Cauchy sequence in X converges to a point in X.*

Definition 1.35. *A complete inner product space is called a **Hilbert space**.*

[35] Gram–Schmidt is one way to calculate the QR decomposition, which reduces ordinary least squares to the exact solution of a triangular system of equations. R and Matlab use the QR decomposition to calculate OLS because it is more stable than the traditional formula when $X'X$ is almost singular.

$\mathbb{R}^n$, l_2, and $L_2[a,b]$ are all Hilbert spaces.

Theorem 1.11. *Classical projection theorem*

Let V be a Hilbert space and let W be a closed subspace[36] *of V. For each $v \in V$, there exists a unique $\widehat{v} \in W$ such that $\|v - \widehat{v}\| \leq \|v - w\|$ for all $w \in W$.*

As before, a necessary and sufficient condition for $\widehat{v}$ to be the unique minimizing vector is that $v - \widehat{v}$ is orthogonal to W.

We call $\widehat{v}$ the **orthogonal projection** of v onto W.

Exercise 1.3. *Let $\{e_1, \ldots, e_n\}$ be an orthonormal basis of W. Show that the orthogonal projection of v onto W is $\sum_{k=1}^{n} \langle v, e_k \rangle e_k$.*

The projection theorem can be framed in another way, which might make the connection to linear regression easier to see.

Definition 1.36. *Let V be an inner product space, and let W be a subspace of V. The **orthogonal complement** of W, denoted $W^\perp$, is the set of vectors orthogonal to W.*

Proposition 1.6. *From the projection theorem: for any Hilbert space V and any closed subspace W of V, any $v \in V$ can be written uniquely in the form $v = \widehat{v} + \epsilon$, where $\widehat{v} \in W$ and $\epsilon \in W^\perp$.*

Matrix Decompositions

Singular value decomposition

The singular value decomposition is used in proofs, and can also be helpful in computations.[37] The idea is that every linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^m$ does the following three things, in order:

1. A rotation and/or reflection in $\mathbb{R}^n$.
2. A scaling along each basis vector in $\mathbb{R}^n$, with the output in $\mathbb{R}^m$.
3. A rotation and/or reflection, now in $\mathbb{R}^m$.

Consider the unit sphere in $\mathbb{R}^n$. Its image under any linear transformation is a hyperellipse. The lengths of the ellipse's axes capture the scaling in the second step, and are called singular values.

Theorem 1.12. *Singular value decomposition (SVD)*

Let $A \in \mathbb{R}^{m \times n}$. Then there exist an $m \times m$ orthogonal matrix U and an $n \times n$ orthogonal matrix V such that

$$A = U \Sigma V'$$

where Σ is a diagonal matrix with $p = \min\{n, m\}$ diagonal entries.

The diagonal entries of Σ are called **singular values** and labeled $(\sigma_1, \ldots, \sigma_p)$. They are uniquely determined,[38] all nonnegative, and conventionally sorted so that $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$. The columns of U are called **left singular vectors** and the columns of V are called **right singular vectors**.

[36] In the sense that W contains all its limit points.

[37] Calculate it using `svd` in R and Matlab, and `np.linalg.svd` in NumPy.

[38] That is, every SVD of A has the same singular values, though perhaps in a different order.

Theorem 1.13. *Reduced SVD*

Suppose that $m \geq n$. Then there exist an $m \times n$ matrix U_1 with orthonormal columns and an $n \times n$ orthogonal matrix V such that

$$A = U_1 \widehat{\Sigma} V'$$

where $\widehat{\Sigma}$ is an $n \times n$ diagonal matrix.

Note that U_1 just contains the first n columns of U; we could partition $U = \begin{pmatrix} U_1 & U_2 \end{pmatrix}$. Intuitively, we go from the reduced SVD to the full SVD by stuffing U_1 with $m - n$ orthonormal vectors and adding $m - n$ zero rows to the bottom of $\widehat{\Sigma}$.

It follows in the $m \geq n$ case that, letting $\{v_1, \dots, v_n\}$ be the columns of V and letting $\{u_1, \dots, u_n\}$ be the columns of U_1,

$$Av_i = \sigma_i u_i, \quad A' u_i = \sigma_i v_i.$$

An interesting consequence is that $(\sigma_1^2, \dots, \sigma_p^2)$ are the eigenvalues of $A'A$ and AA'.

Cholesky factorization

The Cholesky factorization can be helpful for sampling from multivariate normal distributions.[39]

Theorem 1.14. *Cholesky factorization*

Let $A \in \mathbb{R}^{n \times n}$ be symmetric and positive definite. Then there exists a unique lower triangular matrix $L \in \mathbb{R}^{n \times n}$ with positive diagonal entries such that $A = LL'$.

Remark 1.2. *Matrix square root*

Suppose A is a symmetric positive definite matrix and we are interested in finding B such that $B^2 = A$.[40] *Construct it as follows. Let $A = LL'$ using the Cholesky factorization. Let $L = U\Sigma V'$ using the singular value decomposition. Then take*

$$A^{1/2} = U\Sigma U'.$$

You can check that $A^{1/2}$ is symmetric positive definite, and $A^{1/2}A^{1/2} = A$.

Matrix Stacking and the Kronecker Product

When working with matrices as data objects, it is often convenient to move around entries of the matrix to support calculations. The vec operator and the Kronecker product can be useful toward this end.[41]

Definition 1.37. *Let $B \in \mathbb{R}^{m_1 \times n_1}$ with entries b_{ij} and $C \in \mathbb{R}^{m_2 \times n_2}$. Their* ***Kronecker product*** *(or* ***Kronecker tensor product****) $B \otimes C$ is the $m_1 m_2 \times n_1 n_2$ matrix given in the following block matrix form:*

$$B \otimes C = \begin{pmatrix} b_{11}C & \cdots & b_{1n_1}C \\ \vdots & \ddots & \vdots \\ b_{m_1 1}C & \cdots & b_{m_1 n_1}C \end{pmatrix}.$$

[39] Calculate it using `chol` in R and Matlab, and `np.linalg.cholesky` in NumPy.

[40] This is easy if A is diagonal.

[41] Calculate the Kronecker product using `kronecker` in R, `kron` in Matlab, and `np.kron` in NumPy.

The Kronecker product has the following useful properties:

- $(B \otimes C)' = B' \otimes C'$.
- $(B \otimes C)(D \otimes F) = BD \otimes CF$, if the matrices are conformable.[42]
- $(B \otimes C)^{-1} = B^{-1} \otimes C^{-1}$.

When is this useful? The operation of matrix stacking gives a hint.

Definition 1.38. *Let $A \in \mathbb{R}^{m \times n}$ and denote its columns by $v_1, \dots, v_n$. The* ***vec operator*** $\mathrm{vec}(A)$ *maps A to the $nm \times 1$ column vector generated by stacking its columns:*

$$\mathrm{vec}(A) = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}.$$

Proposition 1.7. *If $B \in \mathbb{R}^{m_1 \times n_1}$, $C \in \mathbb{R}^{m_2 \times n_2}$, and $X \in \mathbb{R}^{n_1 \times m_2}$, then*

$$Y = CXB' \iff \mathrm{vec}(Y) = (B \otimes C)\ \mathrm{vec}(X).$$

[42] That is, if matrix multiplication makes sense here.

References

Golub, G. and C. Van Loan. (2013). *Matrix Computations*, 4th ed.

Hoffman, K. and R. Kunze. (1971). *Linear Algebra*, 2nd ed.

Luenberger, D. (1969). *Optimization by Vector Space Methods*.

Chapter 2. Probability

These notes provide a brief introduction to probability. It is intended to provide a simple, very high-level framework for thinking about many of the tools that will be discussed in 2120.

Contents

Principles of Probability

A RANDOM EXPERIMENT is an experiment whose outcome cannot be predicted beforehand. How do we model a random experiment? There are three key elements: The sample space, the events, and the probability measure. We will describe each of these in turn.[1]

Definition 2.1. *The **sample space** Ω is the set of all possible outcomes of a random experiment. We denote an outcome as $\omega \in \Omega$.*

Definition 2.2. *An **event** A is a subset of the sample space, $A \subseteq \Omega$. Let $\mathcal{A}$ denote the family of all events.*

Example 2.1. *Suppose we survey 10 randomly selected people on their employment status and count how many are unemployed.*
The sample space Ω consists of all possible counts of unemployed.

$$\Omega = \{0, 1, 2, \ldots, 10\}$$

A is the event that more than 30% of those surveyed are unemployed.

$$A = \{4, 5, 6, \ldots, 10\}$$

Example 2.2. *Suppose we ask a random person what is their income.*

$$\Omega = \mathbb{R}_+$$

A is the event that the person earns between \$30,000 and \$40,000.

$$A = [30,000, 40,000]$$

Remark 2.1. *These two examples differ in a fundamental way. In the first case, the sample space is a finite set of integers. In example 0.2, the sample space is the set of real numbers. The set of real numbers is a lot bigger than set of integers, and so is the corresponding set of events. The foundations of probability theory will require a little bit of mathematical machinery - in particular, some concepts from a field of analysis called measure theory - in order to define the probability of events for both of these examples in a unified way.*

We place additional restrictions on $\mathcal{A}$. These impose sufficient structure that will allow us to consistently define the probabilities of events.

Definition 2.3. *Let Ω be a set and $\mathcal{A} \subseteq 2^{\Omega}$ be a family of its subsets. $\mathcal{A}$ is a **σ-algebra** if and only if it satisfies the following*

1. *$\Omega \in \mathcal{A}$.*
2. *$\mathcal{A}$ is closed under complements: $A \in \mathcal{A}$ implies that $A^C = \Omega - A \in \mathcal{A}$.*
3. *$\mathcal{A}$ is closed under countable union: If $A_n \in \mathcal{A}$ for $n = 1, 2, \ldots$, then $\cup_n A_n \in \mathcal{A}$.*

[1] Take some time to digest this stuff - the definitions might seem overly complex to describe something that ends up being fairly simple. Bear with us, and welcome to a Ph.D. program in economics!

Remark 2.2. *Properties 1 and 2 of a σ-algebra implies that $\emptyset \in \mathcal{A}$. Properties 2 and 3 imply that $\mathcal{A}$ is closed under countable intersection by DeMorgan's Law. That is, if $A_n \in \mathcal{A}$ for $n = 1, 2, \ldots$, then $\cap_n A_n \in \mathcal{A}$.*

We assume that $\mathcal{A}$, the family of all events on Ω, is a σ-algebra. We say that $(\Omega, \mathcal{A})$ is *measurable space* and that $A \in \mathcal{A}$ is *measurable* with respect to the σ-algebra $\mathcal{A}$.

Definition 2.4. *Let $(\Omega, \mathcal{A})$ be a measurable space. A* ***measure*** *is a function, $\mu : \mathcal{A} \to \mathbb{R}$ such that*

1. *$\mu(\emptyset) = 0$.*
2. *$\mu(A) \geq 0$ for all $A \in \mathcal{A}$.*
3. *If $A_n \in \mathcal{A}$ for $n = 1, 2, \ldots$ with $A_i \cap A_j = \emptyset$ for $i \neq j$, then*

$$\mu(U_n A_n) = \sum_n \mu(A_n).$$

If $\mu(\Omega) < \infty$, we call μ a ***finite measure****. If $\mu(\Omega) = 1$, we call μ a* ***probability measure****. We denote a probability measure as $P : \mathcal{A} \to [0, 1]$.*

Definition 2.5. *A triple $(\Omega, \mathcal{A}, \mu)$ where Ω is a set, $\mathcal{A}$ is a σ-algebra and μ is a measure on $\mathcal{A}$ is a* ***measure space****. If μ is a probability measure, it is* ***probability space****.*

We've now defined all components needed to model a random experiments. A random experiment is characterized by its probability space, $(\Omega, \mathcal{A}, P)$. With the definition of a probability space that we laid out above, we can prove all of the usual probability facts.

Proposition 2.1. *Consider a probability space $(\Omega, \mathcal{A}, P)$. The following hold:*

1. *For all $A \in \mathcal{A}$, $P(A^C) = 1 - P(A)$.*
2. *$P(\Omega) = 1$.*
3. *If $A_1, A_2 \in \mathcal{A}$ with $A_1 \subseteq A_2$, then $P(A_1) \leq P(A_2)$.*
4. *For all $A \in \mathcal{A}$, $0 \leq P(A) \leq P(1)$.*
5. *If $A_1, A_2 \in \mathcal{A}$, then*

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2).$$

Exercise 2.1. *Prove these properties from the definition of a probability space.*

Conditional Probability

Conditional probability gives us a way to model the outcome of a random experiment conditional on some partial information. For instance, given a random experiment and the information that event B has occurred, what is the probability that the outcome also belongs to event A? To do so, we define a new probability measure on Ω.

Definition 2.6. *Let $A, B \in \mathcal{A}$ with $P(B) > 0$. The* ***conditional probability of*** *A* ***given*** *B is*

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

$P(A|B)$ is a probability measure.

Remark 2.3. *We can think about $P(A|B)$ as part of a new probability space with $\Omega = B$ and $P(S) = P(S|B)$ for $S \subseteq B$.*

Remark 2.4. *Because the conditional probability is a probability measure, all of the usual properties of probability measures in Proposition 0.1 apply.*

The definition of conditional probability implies the following useful formula. We have that

$$P(A \cap B) = P(A|B)P(B).$$

We next list several important results about conditional probabilities.

Theorem 2.1. *The multiplication rule*

$$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1)\dots P(A_n| \cap_{i=1}^{n-1} A_i)$$

Proof. This follows via repeated application of the definition of conditional probability. □

Theorem 2.2. *Law of total probability*

Consider K disjoint events C_k that partition Ω. That is, $C_i \cap C_j = \emptyset$ for all $i \neq j$ and $\cup_{i=1}^{K} C_i = \Omega$. Let C be some event.

$$P(C) = \sum_{i=1}^{K} P(C|C_i)P(C_i)$$

Proof. We have that

$$\begin{aligned} C &= C \cap \Omega \\ &= C \cap (\cup_{i=1}^{K} C_i) \\ &= (C \cap C_1) \cup \dots \cup (C \cap C_K) \end{aligned}$$

It follows that

$$P(C) = \sum_{i=1}^{k} P(C \cap C_i)$$

and the result follows from the definition of conditional probability. □

Theorem 2.3. *Bayes' Rule*

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

Exercise 2.2. *Prove Bayes' Rule from the results presented.*

Example 2.3. *Suppose you survey 2 randomly selected individuals. What is the probability that both are female given that at least one is female? Assume that all outcomes are equally likely.*

Solution. The sample space is $\Omega = \{MM, MF, FM, FF\}$. The conditioning event is $B = \{MF, FM, FF\}$ and $A = \{FF\}$. Therefore,

$$P(A|B) = \frac{P(\{FF\})}{P(\{MF, FM, FF\})} = 1/3.$$ ■

As mentioned, we use conditioning to describe the partial information that an event B gives about another event A. What if B provides no information about A?

Definition 2.7. *Two events A, B are* ***independent*** *if*

$$P(A|B) = P(A).$$

Equivalently, they are ***independent*** *if*

$$P(B|A) = P(B)$$

or

$$P(A \cap B) = P(A)P(B).$$

Remark 2.5. *If events A, B are independent, then so are A^C, B, A, B^C and A^C, B^C.*

We can extend the definition of independence to collections of events.

Definition 2.8. *Let $E_1, \ldots, E_n$ be events. $E_1, \ldots, E_n$ are* ***jointly independent*** *if for any $i_1, \ldots, i_k$,*

$$P(E_{i_1}|E_{i_2} \cap \ldots \cap E_{i_k}) = E_{i_1}.$$

Moreover, since conditional probabilities are probability measures, we can define independence with respect to a conditional probability.

Definition 2.9. *Given an event C, events A, B are* ***conditionally independent*** *if*

$$P(A \cap B|C) = P(A|C)P(B|C).$$

Equivalently, A, B are ***independent conditional on*** *C if*

$$P(A|B \cap C) = P(A|C).$$

Random Variables

Borel σ-algebra

Earlier in these notes, we defined a σ-algebra. This was a collection of sets that satisfied some additional restrictions that helped us consistently define the probability of each set. A particularly important σ-algebra is called the **Borel σ-algebra**. This is a σ-algebra over the real line.

Definition 2.10. *Let $\Omega = \mathbb{R}$. Let $\mathcal{A}$ be the collection of all open intervals. The smallest σ-algebra containing all open sets is the* ***Borel σ-algebra****. It is typically denoted as $\mathcal{B}$.*

Note that the Borel σ-algebra contains all closed intervals as well and could have been equivalently defined as the smallest σ-algebra that contains all closed sets. Moreover, we can extend the Borel σ-algebra to higher dimensions: it is the smallest σ-algebra that contains the open balls. The Borel σ-algebra will be useful later on in this section.

Measurable functions

A measurable function is a function that maps from one measure space to another. Measurable functions are useful because for a given set of values in the function's range, we can measure the subset of the function's domain upon which these values occur.

Definition 2.11. *Let $(\Omega, \mathcal{A}, \mu)$ and $(\Omega', \mathcal{A}', \mu')$ be two measure spaces. Let $f : \Omega \to \Omega'$ be a function. f is* ***measurable*** *if and only if $f^{-1}(A') \in \mathcal{A}$ for all $A' \in \mathcal{A}'$.*

That is, $\mu'(f^{-1}(A'))$ is well-defined for a measurable function f. A particularly useful case occurs when

$$(\Omega', \mathcal{A}', \mu') = (\mathcal{R}, \mathcal{B}, \lambda)$$

where λ is the Lebesgue measure. That is, f is a real-valued function. We say that f is **μ-measurable** if and only if

$$f^{-1}((-\infty, c)) = \{\omega \in \Omega : f(\omega) < c\} \in \mathcal{A} \quad \forall c \in \mathbb{R}.$$

We could also state this definition in terms of $>, \leq$ or $\geq$. With these definitions, we are now ready to define a random variable

Random variables

Consider a probability space $(\Omega, \mathcal{A}, P)$. A **random variable** is simply a measurable function from the sample space Ω to the real-line.

Definition 2.12. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X : \Omega \to \mathbb{R}$ is a function. X is a* ***random variable*** *if and only if X is P-measurable. That is, $X^{-1}(B) \in \mathcal{A}$ for all $B \in \mathcal{B}$ where $\mathcal{B}$ is the Borel σ-algebra.*

Definition 2.13. *Let X be a random variable. The* ***cumulative distribution function*** *(cdf) $F : \mathbb{R} \to [0, 1]$ of X is defined as*

$$F_X(x) = P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

For simplicity, we often write

$$F_X(x) = P(X \leq x).$$

Note that $(\mathbb{R}, \mathbb{B}, F_X)$ form a probability space. The cumulative distribution function F_X has the following properties:

1. *For $x_1 \leq x_2$,*
$$F_X(x_2) - F_X(x_1) = P(x_1 < X < x_2).$$
2. $\lim_{x\to-\infty} F_X(x) = 0$, $\lim_{x\to\infty} F_X(x) = 1$.
3. *$F_X(x)$ is non-decreasing.*
4. *$F_X(x)$ is right-continuous:* $\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$.

Remark 2.6. *The* ***quantiles*** *of a random variable X are given by the inverse of its cumulative distribution function. Generally, the* ***quantile function*** *is*

$$Q(u) = \inf\{x : F_X(x) \geq u\}.$$

If F_X is invertible, then

$$Q(u) = F_X^{-1}(u).$$

Remark 2.7. *For any function F that satisfies the properties of a cdf listed above, we can construct a random variable whose cdf is F. Let U be uniformly distributed on $[0, 1]$. That is, $F_U(u) = u$ for all $u \in [0, 1]$. Define $Y = Q(U)$, where Q is the quantile function associated with F. In the case, where F is invertible, we have*

$$F_Y(y) = P(F^{-1}(U) \le y) = P(U \le F(y)) = F(y).$$

Discrete random variables

If F_X is constant except at a countable number of points (i.e. F_X is a step function), then we say that X is a **discrete random variable**. The size of the jump at x_i

$$p_i = F_X(x_i) - \lim_{x \to x_i^-} F_X(x)$$

is the probability that X takes on the value x_i. That is,

$$P(X = x_i) = p_i.$$

The **probability mass function** (pmf) of X is defined as

$$f_X(x) = \begin{cases} p_i & \text{if} \quad x = x_i, \quad i = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}.$$

It follows that we can write

$$P(x_1 < X \le x_2) = \sum_{x_1 < x \le x_2} f_X(x).$$

Continuous random variables

If F_X can be written as

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$$

where $f_X(x)$ satisfies

$$\begin{gathered} f_X(x) \ge 0 \quad \forall x \in \mathbb{R} \\ \int_{-\infty}^{\infty} f_X(t)\, dt = 1, \end{gathered}$$

we say that X is a **continuous random variable**. By the fundamental theorem of calculus, at the points where f_X is continuous,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

We call $f_X(x)$ the **probability density function** (pdf) of X. We call

$$S_X = \{x : f_X(x) > 0\}$$

the **support** of X.

Note that for $x_2 \ge x_1$,

$$\begin{aligned} P(x_1 < X \le x_2) &= F_X(x_2) - F_X(x_1) \\ &= \int_{x_1}^{x_2} f_X(t)\, dt \end{aligned}$$

and that

$$P(X = x) = 0$$

for a continuous random variable.

Remark 2.8. *Do not interpret the pdf of a continuous random variable as expressing a probability: $f_X(x) \neq P(X = x)$. The proper interpretation is that $f_X(x)$ expresses the probability that X falls in some small interval $(x, x + \Delta x)$. That is,*

$$P(X \in (x, x + \Delta x)) \approx f(x)\Delta x.$$

Joint distributions

Let X, Y be two scalar random variables. A **random vector** (X, Y) is a mapping from Ω to $\mathbb{R}^2$.[2] The **joint cumulative distribution function** of X, Y is

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x, Y \leq y) \\ &= P(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq y\}). \end{aligned}$$

We say that (X, Y) is a **discrete random vector** if

$$F_{X,Y}(x, y) = \sum_{u \leq x} \sum_{v \leq y} f_{X,Y}(u, v),$$

where $f_{X,Y}(x, y) = P(X = x, Y = y)$ is the **joint probability mass function** of (X, Y). We say that (X, Y) is a **continuous random vector** if

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) \, dv \, du,$$

where $f_{X,Y}(x, y)$ is the **joint probability density function** of (X, Y). As in the univariate case,

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

at the points of continuity of $F_{X,Y}$. From the joint cdf of (X, Y), we can recover the **marginal cdfs**. We have that

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(X \leq x, Y \leq \infty) \\ &= \lim_{y \to \infty} F_{X,Y}(x, y). \end{aligned}$$

We can also recover the **marginal pdfs** from the joint pdf using

$$f_X(x) = \sum_y f_{X,Y}(x, y) \quad \text{if discrete,}$$

and

$$f_X(x) = \int_{S_Y} f_{X,Y}(x, y) \, dy \quad \text{if continuous.}$$

Consider the discrete case. Let x be such that $f_X(x) > 0$. Then, the **conditional pmf of Y given $X = x$** is

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

This satisfies the following two properties:

$$\begin{aligned} f_{Y|X}(y|x) &\geq 0 \\ \sum_y f_{Y|X}(y|x) &= 1. \end{aligned}$$

[2] We can extend the formal definitions for random variables to random vectors. (X, Y) is a random vector if and only if (X, Y) is P-measurable. That is, $(X, Y)^{-1}(B) \in A$ for all $B \in \mathbb{B}$, where $\mathbb{B}$ is now the Borel σ-algebra on $\mathbb{R}^2$.

That is, $f_{Y|X}(y|x)$ is a well-defined pmf for a discrete random variable. The **conditional cdf** of Y given $X = x$ is then

$$F_{Y|X}(y|x) = P(Y \leq y|X = x) = \sum_{v \leq y} f_{Y|X}(v|x).$$

Next, consider the continuous case. It is analogous. For any x such that $f_X(x) > 0$, the **conditional pdf** of Y given $X = x$ is

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

Provided that $f_X(x) > 0$, this is a well-defined pdf for a continuous random variable. The **conditional cdf** is

$$F_{Y|X}(y|x) = \int_{-\infty}^{y} f_{Y|X}(v|x)\, dv.$$

Remark 2.9. *The conditional pmf for two discrete random variables can be interpreted as a probability. That is, for the discrete random vector (X, Y),*

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

However, this is not true for continuous random variables because if X is continuous, $P(X = x) = 0$. Instead, think about it as

$$\begin{aligned} f_{Y|X}(x,y) &= P(X \in dx|Y = y) \\ &= \lim_{\Delta y \to 0} P(X \in dx|y \leq Y \leq y + \Delta y). \end{aligned}$$

Finally, we extend the definition of independence to random variables. The random variables X, Y are **independent** if $F_{Y|X}(y|x) = F_Y(y)$ or equivalently, if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$. We can also define this in terms of the densities. X, Y are independent if $f_{Y|X}(y|x) = f_Y(y)$ or equivalently, if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

All of these definitions and results extend to n-dimensional random variables in a straightforward manner.

Transformations of random variables

Let X be a random variable with cdf F_X. Define the random variable $Y = h(X)$, where h is a one-to-one function whose inverse h^{-1} exists. What is the distribution of Y?

First, suppose that X is discrete and takes on values $x_1, \ldots, x_n$. Y is also discrete and takes on the values

$$y_i = h(x_i), \quad \text{for} \quad i = 1, \ldots, n.$$

We have that the pmf of Y is given by

$$\begin{aligned} P(Y = y_i) &= P(X = h^{-1}(x_i)) \\ f_Y(y) &= f_X(h^{-1}(y_i)). \end{aligned}$$

Next, suppose that X is continuous. Consider the case where h is increasing. We have that

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(X \leq h^{-1}(y)) = F_X(h^{-1}(y)). \end{aligned}$$

It follows directly that

$$\begin{aligned} f_Y(y) &= \frac{dF_Y(y)}{dy} \\ &= f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}. \end{aligned}$$

In the case where h is decreasing, we can analogously show that

$$f_Y(y) = -f_X(h^{-1}(y)\frac{dh^{-1}(y)}{dy}.$$

Combining these two cases, we have that, in general,

$$f_Y(y) = f_X(h^{-1}(y))\left|\frac{dh^{-1}(y)}{dy}\right|.$$

Example 2.4. *Suppose $X \sim U[0,1]$ and $Y = X^2$. Over the support of X, this is a one-to-one transformation. We have that*

$$X = \sqrt{Y}, \quad dX/dY = \frac{1}{2}y^{-1/2}.$$

Applying the formula above, we have that $S_Y = [0,1]$ and

$$f_Y(y) = \frac{1}{2}y^{-1/2}.$$

This can be extended to the multivariate case. Let X be a random vector and as before, define $Y = h(X)$. You can show that

$$f_Y(y) = f_X(h^{-1}(x))\,|J|$$

where $|J|$ is the absolute value of the determinant of the Jacobian matrix of the inverse transformation. That is, $|J|$ is the absolute value of the determinant of the matrix whose i,j-th entry is $\partial x_i/\partial y_j$.

Expectations

Suppose X is a discrete random variable. Its **expectation** or **expected value** is defined as

$$E[X] = \sum_x x f_X(x).$$

if $\sum_x |x| f_X(x) < \infty$. Otherwise, its expectation is said to not exist. Suppose X is a continuous random variable. Its expectation is defined as

$$E[X] = \int_{S_X} x f_X(x)\,dx$$

if $\int_{S_X} |x| f_X(x)\,dx < \infty$. Otherwise, its expectation is said to not exist.[3] We can also define the expectation of functions of random variables. Let $g : \mathbb{R} \to \mathbb{R}$. Then, if X is discrete,

$$E[g(X)] = \sum_x g(x) f_X(x)$$

and if X is continuous, then

$$E[g(X)] = \int_{S_X} g(x) f_X(x)\,dx.$$

The following are useful properties of the expectation operator. Suppose $a, b \in \mathbb{R}$ and $g_1(\cdot), g_2(\cdot)$ are real-valued functions.

[3] Formally, the expectation is defined using the Lebesgue-Stieltjes integral.

1. $E[a] = a$.
2. $E[ag_1(X)] = aE[g_1(X)]$.
3. $E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$.

These properties together imply that the expectation is a *linear operator*.

We can use the expectation operator to express probabilities. An **indicator function** $1(A)$ is a function that is equal to one if condition A is true and zero otherwise. For example, if X is a random variable, then

$$1(X \leq x) = \begin{cases} 1 & \text{if} \quad X \leq x \\ 0 & \text{otherwise} \end{cases}.$$

Note that (for the continuous case)

$$\begin{aligned} E[1(X \leq x)] &= \int_{-\infty}^{\infty} 1(X \leq x) f_X(x)\, dx \\ &= \int_{-\infty}^{x} f_X(x)\, dx \\ &= F_X(x) = P(X \leq x). \end{aligned}$$

More generally, if $A_X \subseteq \mathbb{R}$, we have that

$$E[1(X \in A_X)] = P(X \in A_X).$$

This is a very useful result.

Suppose X, Y are random variables with joint density $f_{X,Y}(x, y)$. Let $g(x, y) : \mathbb{R}^2 \to \mathbb{R}$. We have that

$$E[g(X, Y)] = \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dy\, dx.$$

Note that by linearity of the expectation, for $a, b \in \mathbb{R}$,

$$E[aX + bY] = aE[X] + bE[Y].$$

Finally, if X, Y are independent, then for any functions $h_1(\cdot), h_2(\cdot)$,

$$E[h_1(X)h_2(Y)] = E[h_1(X)]E[h_2(Y)].$$

All of these results generalize directly to higher dimensions.

Conditional expectations

Given a pair of random variables (X, Y) with a joint density $f_{X,Y}(x, y)$, we can define the **conditional expectation** of Y given $X = x$ as

$$E[Y|X = x] = \int_{S_Y} y f_{Y|X}(y|x)\, dy.$$

Note that this is a function of x. It is sometimes denote $\mu_Y(x)$ and called the **regression function**. In particular, this means we can view this a random function $E[Y|X]$. The following theorem is extremely useful.[4]

Theorem 2.4. *The law of iterated expectations*[5]

$$E_Y[Y] = E_X E_{Y|X}[Y],$$

where E_X denotes the expectation taken with respect to the marginal density of X and $E_{Y|X}$ denotes the expectation taken with respect to the conditional density of Y given X.

[4] This might be the most important thing we cover in math camp!

[5] This is also called the Tower Property.

Proof. We have that

$$\begin{aligned} E_X E_{Y|X}[Y] &= \int \left(\int y f_{Y|X}(y)\, dy \right) f_X(x)\, dx \\ &= \int \int y f_{Y|X}(y) f_X(x)\, dy\, dx \\ &= \int y \left(\int f_{X,Y}(x,y)\, dx \right) dy \\ &= \int y f_Y(y)\, dy = E[Y]. \end{aligned}$$

□

What are some ways to interpret the conditional expectation? We provided a formal definition but we also want to provide some intuition. First, the conditional expectation is the solution to an *optimal forecasting* problem. Suppose you wish to forecast the value of a random variable Y. That is, we wish to pick $h \in \mathbb{R}$ that minimizes the expected mean-square error

$$E[(Y-h)^2] = \int (y-h)^2 f_Y(y)\, dy.$$

The first order condition is

$$\int y f_Y(y)\, dy = \int h f_Y(y)\, dy \implies h^* = E[Y].$$

That is, the optimal prediction of Y is $E[Y]$.[6] Now, suppose that we observe another random variable X and see that $X = x$. We wish to forecast Y as a function of x. That is, we wish to minimize

$$E[(Y - h(X))^2].$$

Note that we can always write any function of X as

$$h(x) = \mu_Y(x) + g(x)$$

by defining $g(x) = h(x) - \mu_Y(x)$. So choosing h to minimize expected mean-square error is equivalent to choosing g. We can then write

$$(Y - h(X))^2 = (Y - \mu_Y(X))^2 - 2g(X)(Y - \mu_Y(x)) + g(X)^2.$$

I claim that[7]

$$E_{Y|X}[g(X)(Y - \mu_Y(x))] = 0$$

and so,

$$E[(Y - h(X))^2] = E[(Y - \mu_Y(X))^2 + g(X)^2].$$

It then follows immediately that expected mean-squared error is minimized with $g(x) = 0$ and so,

$$h^*(x) = \mu_Y(x).$$

That is, the conditional expectation of Y given X is the optimal predictor of Y given X.[8]

Second, we can interpret the conditional expectation of Y given X as the orthogonal projection of Y onto the space of functions of the random variable X, i.e., L^2 space. Since this interpretation of the conditional expectation is the focus of the first several lectures of Econ 2120, I will not cover it here.

[6] Optimal with respect to expected mean-square error. If we changed the objective function to expected mean-absolute error, $E[|Y - h|]$, the solution is the median of Y, $h^* = median(Y)$.

[7] Can you show these steps?

[8] And with that, you have learned a good chunk of machine learning. I am not joking.

Moments and moment generating functions (MGFs)

Consider a random variable X. The **k-th moment of** X is defined as $E[X^k]$. The first moment of X is its **mean**, $E[X]$. The **k-th centered moment of** X is $E[(X - E[X])^k]$. The second centered moment of X is its **variance**, $V(X) = E[(X - E[X])^2]$. The **standard deviation** of X is $\sqrt{V(X)}$.

Remark 2.10. *Suppose X has mean μ_X and variance σ_X^2. Let $a, b \in \mathbb{R}$ and define $Y = a + bX$. Then,*

$$\mu_Y = a + b\mu_X, \quad \sigma_Y^2 = b^2\sigma_X^2.$$

Definition 2.14. *The **moment generating function** (MGF) of a random variable X is defined as*

$$\mu_X(t) = E[e^{tX}] = \int e^{tx} f_X(x)\, dx.$$

The MGF of X is useful because it allows us to easily compute all of the moments of a random variable. Note that

$$\mu_X'(t) = \int x e^{tx} f_X(x)\, dx, \quad \mu_X'(0) = \int x f_X(x)\, dx = E[X],$$
$$\mu_X''(t) = \int x^2 e^{tx} f_X(x)\, dx, \quad \mu_X''(0) = \int x^2 f_X(x)\, dx = E[X^2].$$

In general, we can show that

$$\mu_X^{(j)}(0) = E[X^j] \quad \text{for} \quad j = 1, 2, \ldots$$

Moreover, the MGF of a random variable completely characterizes the distribution of a random variable. If X, Y are two random variables with the same MGF, then they have the same distribution.

Remark 2.11. *The MGF may not always exist for a random variable. For example, e^{tX} may blow up for large realizations of X. However, the **characteristic function** of X is guaranteed to exist. It is defined as*

$$E[e^{itx}], \quad i = \sqrt{-1}.$$

The characteristic function is guaranteed to exist and it completely characterizes the distribution of X.

Moments for random vectors

Suppose X, Y are two random variables with joint density $f_{X,Y}(x, y)$. The **covariance** between X, Y is defined as

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

The covariance is a linear operator. That is,

$$Cov(X, aY + bW) = aCov(X, Y) + bCov(X, W).$$

Moreover, suppose $Z = aX + bY$ for $a, b \in \mathbb{R}$. Then,

$$V(Z) = a^2V(X) + b^2V(Y) + 2ab\, Cov(X, Y).$$

Now suppose that X is an n-dimensional random vector with $X = (X_1, \dots, X_n)$. Its **mean vector** is

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

and its **covariance matrix** is

$$V(X) = \Sigma$$

where Σ is an $n \times n$ matrix whose ij-th entry is $\Sigma_{ij} = Cov(X_i, X_j)$. Σ is a positive semi-definite matrix. Why? Let $\alpha \in \mathbb{R}^n$ and define $Y = \alpha' X$. Then,

$$V(Y) = \alpha' \Sigma \alpha \geq 0.$$

This must hold for all $\alpha \in \mathbb{R}^n$.

Useful Probability Distributions

Bernoulli distribution

X is a discrete random variable that can only take on two values: $0, 1$. We write $f_X(1) = p$, $f_X(0) = 1 - p$ and so,

$$f_X(x) = p^x (1-p)^{1-x}.$$

Note that

$$\begin{aligned} &E[X^k] = p, \quad k \geq 1 \\ &V(X) = p(1-p), \\ &\mu_X(t) = (1-p) + pe^t. \end{aligned}$$

We say that X has a **Bernoulli distribution**.

Binomial distribution

Suppose that X_i for $i = 1, \dots, n$ are i.i.d. Bernoulli random variables with $P(X_i = 1) = p$. Define $X = \sum_{i=1}^n X_i$. We say that X follows a **binomial distribution** with parameters n and p. X takes on values $1, 2, \dots, n$. Its pmf is

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

and

$$E[X] = np, \quad V(X) = np(1-p).$$

Poisson distribution

Suppose that X is a discrete random variable and takes on values $0, 1, 2, 3, \dots$. Its pmf is

$$f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \lambda > 0.$$

We say that X is a **Poisson** random variable with parameter $\lambda > 0$. Note that

$$E[X] = \lambda, \quad V(X) = \lambda.$$

Poisson random variables are typically used to model the number of discrete "successes" that occur over a time period.

Remark 2.12. *Note that if X_n is binomially distributed with parameters $n, p = \lambda/n$, then $X_n \xrightarrow{d} X$, where X is a Poisson random variable.*[9]

Uniform distribution

Suppose that X is a continuous random variable with $f_X(x) = \frac{1}{b-a}$ for $x \in [a, b]$ and 0 otherwise. We say that X is **uniformly distributed on [a, b]** and write $X \sim U[a, b]$.

Univariate normal distribution

Suppose Z is continuously distributed with support over $\mathbb{R}$. We say that Z follows a **standard normal distribution**, $Z \sim N(0, 1)$, if

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

Note that $E[Z] = 0$ and $V(Z) = 1$. We say that $X \sim N(\mu, \sigma^2)$ if

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}.$$

Note that $E[X] = \mu$, $V(X) = \sigma^2$ and $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$.

The MGF of a standard normal random variable is incredibly useful.[10] It is worth memorizing. If $Z \sim N(0, 1)$, then

$$M_Z(t) = e^{\frac{1}{2}t^2}.$$

Why? Here's the calculation:[11]

$$\begin{aligned}
M_Z(t) &= E[e^{tZ}] \\
&= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\, dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz-\frac{1}{2}z^2}\, dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2tz)}\, dz \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2-2tz+t^2)}\, dz \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}\, dz \\
&= e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w)^2}\, dz, \quad w = z - t \\
&= e^{\frac{1}{2}t^2}.
\end{aligned}$$

[9] The notation $X_n \xrightarrow{d} X$ means that the sequence of random variables X_n converges in distribution to the random variable X. See the note on asymptotics for the formal definition. Intuitively, it means that as n gets large, X_n behaves as if it were distributed like X.

[10] For example, you'll run into it all the time in macro/finance.

[11] It's straightforward provided you remember how to complete the square.

We can use this to derive the MGF for $X \sim N(\mu, \sigma^2)$. We have that

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= E[e^{t(\mu+\sigma Z)}] \\ &= e^{t\mu} E[e^{t\sigma Z}] \\ &= e^{t\mu} M_Z(t\sigma) \\ &= e^{\mu t + \frac{1}{2}\sigma^2 t^2}. \end{aligned}$$

Therefore, we have that

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}.$$

Chi-squared distribution

Let $Z_i \sim N(0,1)$ i.i.d. for $i = 1, \ldots, n$. Let

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

We say that X is a **chi-squared** random variable with n **degrees of freedom** and write $X \sim \chi_n^2$. It follows immediately that

$$E[X] = n, \quad V(X) = 2n.$$

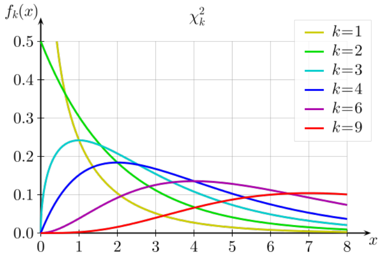

Figure 2: Density of χ^2 as degree of freedom varies. (Source: Wikipedia)

F-distribution

Let $Y_1 \sim \chi_k^2$, $Y_2 \sim \chi_l^2$ with $Y_1 \perp Y_2$. Define

$$Q = \frac{Y_1/k}{Y_2/l}.$$

We say that Q follows an **F-distribution** with k, l degrees of freedom. We write $Q \sim F_{k,l}$.

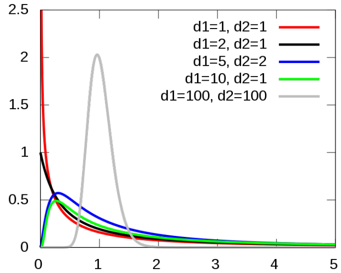

Figure 3: Density of $F_{k,l}$ as degrees of freedom vary. (Source: Wikipedia)

Student t-distribution

Let $Z \sim N(0,1)$ and $Y \sim \chi_n^2$ with $Z \perp Y$. Define

$$T = Z/\sqrt{Y/n}.$$

We say that T is **student t-distributed** with n **degrees of freedom** and write $T \sim t_n$. We have that

$$E[T] = 0$$

$$V(T) = \begin{cases} \frac{n}{n-2}, & \text{if} \quad n > 2, \\ \infty, & \text{if} \quad n = 1, 2 \end{cases}.$$

Remark 2.13. *As $n \to \infty$, $T_n \xrightarrow{d} Z \sim N(0,1)$. This result is the foundation of asymptotic inference in econometrics.*

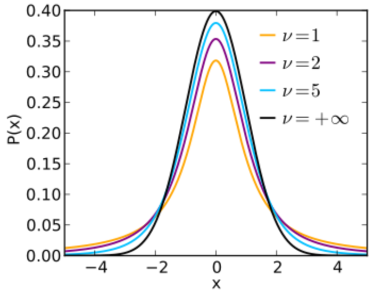

Figure 4: Density of t_n as degree of freedom varies. (Source: Wikipedia)

Exponential distribution

Suppose that X is a continuous random variable with support over $\mathbb{R}_+$. X is **exponentially distributed** with parameter $\lambda > 0$ if

$$f_X(x) = \lambda e^{-\lambda x}.$$

We write $X \sim exp(\lambda)$ and have that

$$E[X] = \frac{1}{\lambda}$$
$$V(X) = \frac{1}{\lambda^2}.$$

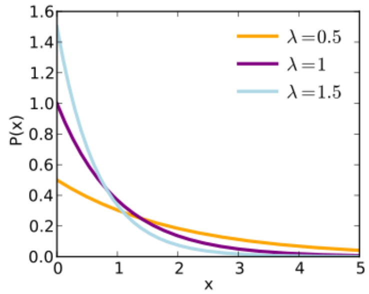

Figure 5: Distribution of $exp(\lambda)$ as λ. (Source: Wikipedia)

Multivariate Normal Distribution

Consider the random vector $Z = (Z_1, \dots, Z_m)'$, where each $Z_i \sim N(0,1)$ i.i.d. The joint density of Z is given by

$$\begin{aligned} f_Z(z) &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} \\ &= (1/2\pi)^{n/2} \exp(-\frac{1}{2}\sum_{i=1}^{m} z_i^2) \\ &= (2\pi)^{n/2} \exp(-\frac{1}{2} z'z) \end{aligned}$$

Moreover, note that $E[Z] = 0$ and $V(Z) = I_m$. Finally, the MGF of Z is given by

$$\begin{aligned} M_Z(t) &= E[e^{t'Z}] \\ &= E[\prod_{i=1}^{m} e^{t_i z_i}] \\ &= \prod_{i=1}^{m} E[e^{t_i z_i}] = e^{\frac{1}{2}t't}. \end{aligned}$$

This is a useful reference point as we develop some results about the multivariate normal distribution.

Definition 2.15. *A m-dimensional random vector X follows a m-**dimensional multivariate normal distribution** if and only if*

$$a'X$$

is normally distributed for all $a \in \mathbb{R}^m$. We write $X \sim N_m(\mu, \Sigma)$, where $E[X] = \mu$ is the m-dimensional mean vector and $V(X) = \Sigma$ is the $m \times m$ dimensional covariance matrix.[12]

Remark 2.14. *If X follows a multivariate normal distribution, then each element X_i follows a univariate normal distribution with mean μ_i and variance Σ_{ii}.*

Remark 2.15. *It is important to have a good handle on the properties of the univariate and multivariate normal distributions. When we use asymptotics to approximate the finite-sample distribution of estimators and test-statistics in econometrics, everything "becomes" normally distributed by the central theorem.*[13]

[12] Typically, the dimension is suppressed in the notation. That is, if X is m-dimensional and follows a multivariate normal distribution, we will typically write $X \sim N(\mu, \Sigma)$. The dimensions of μ, Σ are implied by the context.

[13] Not literally everything but you get the point.

The next two results will allow us to derive the distribution of a multivariate normal. We first derive its MGF.

Proposition 2.2. *Suppose $X \sim N(\mu, \Sigma)$. Then,*

$$M_X(t) = e^{t'\mu + \frac{1}{2}t'\Sigma t}.$$

Proof. Note that $t'X \sim N(t'\mu, t'\Sigma t)$. Therefore,

$$\begin{aligned} M_X(t) &= E[e^{t'X}] \\ &= E[e^Y], \quad Y \sim N(t'\mu, t'\Sigma t) \\ &= M_Y(1) \end{aligned}$$

and the result follows. □

Recall that for a univariate normal distribution, if $X \sim N(\mu, \sigma^2)$, then $Y = aX + b \sim N(a\mu + b, a^2\sigma^2)$. The same property holds for the multivariate normal distribution.

Proposition 2.3. *Suppose $X \sim N_m(\mu, \Sigma)$. Define*

$$Y = AX + b,$$

where $A \in \mathbb{R}^{n \times m}, b \in \mathbb{R}^n$. Then,

$$Y \sim N_n(A\mu + b, A\Sigma A').$$

Proof. For $t \in \mathbb{R}^n$,

$$\begin{aligned} M_Y(t) &= E[e^{t'Y}] \\ &= E[e^{t'(AX+b)}] \\ &= e^{t'b}E[e^{(A't)'X} \\ &= e^{t'b}e^{(A't)'\mu + \frac{1}{2}(A't)'\Sigma(A't)'} \\ &= e^{t'(A\mu+b)+\frac{1}{2}t'(A\Sigma A')t}. \end{aligned}$$

□

We'll now use the two previous results to derive the density of a multivariate normal distribution.

Proposition 2.4. *Suppose $X \sim N(\mu, \Sigma)$ and Σ has full column rank. Then, the density of X is given by*

$$f_X(x) = (2\pi)^{-m/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)).$$

Proof. Let Z be a m-dimensional random vector of i.i.d. standard normal random variables. At the beginning of this section, we derived that $M_Z(t) = e^{\frac{1}{2}t't}$. Therefore, $Z \sim N_m(0, I_m)$. We also derived that the density of Z is

$$f_Z(z) = (2\pi)^{-m/2}e^{-\frac{1}{2}z'z}.$$

Let $X = \mu + \Sigma^{1/2}Z$. We can show that $X \sim N_m(\mu, \Sigma)$. From the multivariate transformation of random variables formula from an earlier section,

$$f_X(x) = |\Sigma|^{-1/2}f_Z(\Sigma^{-1/2}(x-\mu))$$

and the result follows. □

The rest of this section lists additional useful properties of the multivariate normal distribution that will appear from time to time. It's useful to be familiar with them.

Proposition 2.5. *If $X_1 \sim N_m(\mu_1, \Sigma_1), X_2 \sim N_n(\mu_2, \Sigma_2)$ and $X_1 \perp X_2$, then*

$$X = (X_1', X_2')' \sim N_{m+n}(\mu, \Sigma)$$

where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}.$$

Proposition 2.6. *Let $X \sim N_m(\mu, \Sigma)$. Let X_1 be a p-dimensional sub-vector of X with $p < m$. Write*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

and

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then, $X_1 \sim N_p(\mu_1, \Sigma_{11})$.

Proposition 2.7. *Let $X \sim N_m(\mu, \Sigma)$. Partition X into two sub-vectors. That is, write*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

and

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then, $X_1 \perp X_2$ if and only if $\Sigma_{12} = \Sigma_{21} = 0$.

Proposition 2.8. *Let $X \sim N_m(\mu, \Sigma)$. If*

$$Y = AX + b, \quad V = CX + d,$$

where $A, C \in \mathbb{R}^{n \times m}$ and $b, d \in \mathbb{R}^n$, then

$$Cov(Y, V) = A\Sigma C'.$$

Moreover, $Y \perp V$ if and only if

$$A\Sigma C' = 0.$$

Exercise 2.3. *Prove these properties of the multivariate normal distribution.*

Proposition 2.9. *Let $X \sim N_m(\mu, \Sigma)$ with $X = (X_1', X_2')'$, $\mu = (\mu_1', \mu_2')'$ and*

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Provided that Σ_{22} has full rank, the conditional distribution of X_1 given $X_2 = x_2$ is

$$X_1 | X_2 = x_2 \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

Remark 2.16. *What's the intuition of this? We have that*

$$E[X_1|X_2 = x_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2).$$

This formula will look more familiar if everything is one-dimensional. It becomes

$$E[X_1|X_2 = x_2] = E[X_1] + \frac{Cov(X_1, X_2)}{V(X_2)}(x_2 - E[X_2]).$$

Is this starting to look more familiar? Not yet? Ok, let's relabel $Y = X_1, X = X_2$ and re-arrange. Then,

$$E[Y|X = x] = (E[Y] - \frac{Cov(Y, X)}{V(X)}E[X]) + \frac{Cov(Y, X)}{V(X)}x.$$

This is simply the linear regression formula![14] *For a multivariate normal random distribution, conditional expectations are exactly linear. As a result, linear regression exactly returns the conditional expectation function.*

This final result provides the conditional distribution of a multivariate normal distribution. This appears at random points throughout the first year and so, it is useful to keep in your back pocket.

Quadratic forms of normal random vectors

Recall that a **quadratic form** is a quantity of the form $y'Ay$, where A is a symmetric matrix. Suppose that $Z_i \sim N(0, 1)$ i.i.d. for $i = 1, \ldots, n$. We already know that $\sum_{i=1}^{n} Z_i^2 = Z'Z \sim \chi_n^2$.

Proposition 2.10. *If $X \sim N_m(\mu, \Sigma)$ and Σ has full rank, then*

$$(X - \mu)'\Sigma^{-1}(X - \mu) \sim \chi_m^2.$$

Proof. Let $Z = \Sigma^{-1/2}(X - \mu) \sim N_m(0, I_m)$. Then, $Z'Z \sim \chi_m^2$. □

Jensen, Markov and Chebyshev, Oh My!

The following are some useful inequalities that pop up in a variety of contexts in econometrics and other areas of economics. These are especially useful in asymptotics.

Theorem 2.5. *Jensen's inequality*

Let $h(\cdot)$ be a convex function and X be a random variable. Then,

$$E[h(X)] \geq h(E[X]).$$

Proof. Recall that if $h\cdot$ is a convex function, then $\forall x_0$ in its domain, there exists a line through $(x_0, h(x_0))$ such that $h(x)$ never falls below the line. That is, there exists some constant a such that

$$h(x) \geq h(x_0) + a(x - x_0) \quad \forall x.$$

Set $x_0 = E[x]$. It follows that

$$h(X) \geq h(E[X]) + a(x - E[X])$$

holds for all x. Taking expectations, we have that

$$E[h(X)] \geq h(E[X]). \quad \square$$

[14] Set $\beta_0 = E[Y] - \frac{Cov(Y,X)}{V(X)}E[X]$ and $\beta_1 = \frac{Cov(Y,X)}{V(X)}$. Then, $E[Y|X = x] = \beta_0 + \beta_1 x$.

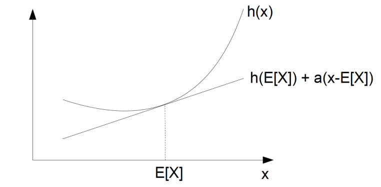

Figure 6: Picture proof of Jensen's inequality.

Remark 2.17. *If $h(\cdot)$ is concave, the opposite inequality holds. That is,*

$$E[h(X)] \leq h(E[X]).$$

The next inequality (Markov's inequality) provides a bound on the tail behavior of a random variable as a function of its expectation.

Theorem 2.6. *Markov's inequality*

Suppose X is a random variable with $X \geq 0$ with $E[X] < \infty$.[15] *Then, for all $M > 0$,*

$$P(X \geq M) \leq \frac{E[X]}{M}.$$

Proof. The proof is straightforward. Because $X \geq 0$,

$$X \geq M1(X \geq M).$$

Taking expectations of both sides, we have that

$$E[X] \geq ME[1(X \geq M) = MP(X \geq M)$$

annd re-arrange to arrive at the result. □

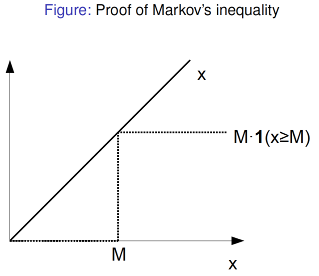

Figure 7: Picture proof of Markov's inequality.

Example 2.5. *Suppose that household income is non-negative. By Markov's inequality, no more than $1/5$ of households can have an income that is greater than five times the average household income.*

The final inequality (Chebyshev's inequality) is a corollary of Markov's inequality. It provides an upper bound on the probability that a random variable falls a certain distance from its expectation.

Theorem 2.7. *Chebyshev's inequality*

Suppose that X is a random variable such that $\sigma^2 = V(X) < \infty$. Then, for all $M > 0$,

$$P(|X - E[X]| > M) \leq \frac{\sigma^2}{M^2}.$$

Proof. Let $Y = (X - E[X])^2$. Apply Markov's inequality to Y and the cutoff M^2 to get

$$P(Y \geq M^2) \leq \frac{E[Y]}{M^2}.$$

Rewrite to get that

$$P(|X - E[X]| \geq M) \leq \frac{\sigma^2}{M^2}.$$

□

Example 2.6. *Chebyshev's inequality is used in a proof of the weak law of large numbers (WLLN).*[16] *For now, WLLN states: as the sample size gets very large, the sample average of a random variable "converges" to the expectation of the random variable. One proof begins by showing that the variance of the sample average converges to zero and then uses Chebyshev's inequality to prove the result.*[17]

[15] $X \geq 0$ for a random variable means that $P(\{\omega : X(\omega) < 0\}) = 0$.

[16] The weak law of large numbers will be introduced in detail later.

[17] If that made no sense, don't worry about it. We'll go through this together.

References

Billingsley, P. (2012). *Probability and Measure*.

Blitzstein, J. and J. Hwang. (2014). *Introduction to Probability*.

Casella, G. and R. Berger. (2001). *Statistical Inference*.

Hogg, R., McKean, J. and A. Craig. (2012). *Introduction to Mathematical Statistics*.

Kolmogorov, A. and Fomin, S. (2012). *Introductory Real Analysis*.

Stokey, N., Lucas, R. and E. Prescott. (1989). *Recursive Methods in Economic Dynamics*.

Chapter 3. Asymptotics

These notes provide a brief introduction to asymptotics. The idea of asymptotic analysis is to ask how a given estimator might behave under 'ideal conditions' - usually (but not always), this means studying the limiting behavior of an estimator as the number of observations grows large. This is an enormously helpful simplifying assumption, and usually allows us to say things about how our estimator behaves under idealized conditions.

Contents

THE WORLD IS VERY COMPLEX and we often do not want to make strict parametric assumptions in our econometric models.[1] Can we still say something about the behavior of our estimators without these strict assumptions? It turns out that we can *in large samples*. We ask the question: How would my estimator behave in very large samples?[2] We then use the limiting behavior of our estimator in infinitely large samples to approximate its behavior in finite samples.

Of course this approach has its advantages and disadvantages. As the sample size gets infinitely large, the behavior of most estimators becomes very simple. In most cases, we can apply some version of the central limit theorem and so, our estimator behaves as if its sampling distribution were normal in large samples. However, this is only an *approximation* for the true, finite-sample distribution of the estimator, which is typically unknown. So it's possible that in finite samples, asymptotic approximations used to construct standard errors (for instance) can be really bad.

In this note, we will summarize the basic tools necessary for asymptotic statistics. A large portion of econometrics revolves around deriving these asymptotic approximations, finding out when these approximations are poor and what to do about it.

Types of Convergence

Recall the definition of convergence for a non-stochastic sequence of real numbers. Let $\{x_n\}$ be a sequence of real numbers. We say

$$\lim_{n \to \infty} x_n = x$$

if for all $\epsilon > 0$, there exists some N such that for all $n > N$, $|x_n - x| < \epsilon$. We want to generalize this to the convergence of random variables. That is, under what conditions does the sequence of random variables $\{X_n\}$ "converge" to another random variable X? There are several notions of stochastic convergence.[3]

Definition 3.1. *The sequence of random variables $\{X_n\}$* ***converges to the random variable** X* ***almost surely*** *if*

$$P(\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\}) = 1.$$

We write

$$X_n \xrightarrow{as} X.$$

Remark 3.1. *What does almost sure convergence mean? For a given outcome ω in the sample space Ω, we can ask whether*

$$\lim_{n \to \infty} X_n(\omega) = X(\omega)$$

holds using the definition of non-stochastic convergence. If the set of outcomes for which this holds has probability one then $X_n \xrightarrow{as} X$.

Definition 3.2. *The sequence of random variables $\{X_n\}$* ***converges to the random variable** X* ***in probability*** *if for all $\epsilon > 0$,*

$$\lim_{n \to \infty} P(|X_n - X| > \epsilon) \to 0.$$

[1] For instance, it's often not realistic in empirical applications to assume that the error term in a linear regression is normally distributed.

[2] As the sample size n goes to infinity.

[3] All of these random variables are defined on the same sample space.

We write

$$X_n \xrightarrow{p} X.$$

Remark 3.2. *What does convergence in probability mean? Fix an $\epsilon > 0$ and compute*

$$P_n(\epsilon) = P(|X_n - X| > \epsilon).$$

This is just a number and so, we can check whether $P_n(\epsilon) \to 0$ using the definition of non-stochastic convergence. If $P_n(\epsilon) \to 0$ for all values $\epsilon > 0$, then $X_n \xrightarrow{p} X$.

Definition 3.3. *The sequence of random variables $\{X_n\}$* ***converges in mean to the random variable*** *X if*

$$\lim_{n\to\infty} E[|X_n - X|] = 0.$$

We write

$$X_n \xrightarrow{m} X.$$

Definition 3.4. *$\{X_n\}$* ***converges in mean-square to*** *X if*

$$\lim_{n\to\infty} E[|X_n - X|^2] = 0.$$

We write

$$X_n \xrightarrow{ms} X.$$

Remark 3.3. *$m_n = E[|X_n - X|]$ is just a number. $X_n \xrightarrow{m} X$ if and only if $m_n \to 0$ using the definition of non-stochastic convergence. Similarly, $ms_n = E[|X_n - X|^2]$ is also just a number and we can think about mean-square convergence in the same way.*

Definition 3.5. *Let $\{X_n\}$ be a sequence of random variables and $F_n(\cdot)$ is the cdf of X_n. Let X be a random variable with cdf $F(\cdot)$. $\{X_n\}$* ***converges in distribution****,* ***weakly converges*** *or* ***converges in law*** *to X if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for all points x at which $F(x)$ is continuous. There are many ways of writing this

$$X_n \xrightarrow{d} X$$
$$X_n \xrightarrow{\mathcal{L}} X$$
$$X_n \implies X.$$

We'll use $X_n \xrightarrow{d} X$.

Remark 3.4. *Convergence in distribution describes the behavior of CDFs. It *does not* mean that realizations of the involved random variables will be close to each other. Recall that $F(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\})$. As a result, $F_n(x) \to F(x)$ does not make any statement about $X_n(\omega)$ getting close to $X(\omega)$ for any $\omega \in \Omega$.*

To illustrate why convergence in distribution is restricted to the continuity points of $F(x)$, consider the following example.

Example 3.1. *Let X_n be a degenerate random variable defined by $X_n = 1/n$ with probability 1 and let X be a degenerate random variable defined by $X = 0$ with probability one. Then, $F_n(x) = 1(x \geq 1/n)$ and $F(x) = 1(x \geq 0)$ with $F_n(0) = 0$ for all n while $F(0) = 1$.*

However, as $n \to \infty$, X_n is getting closer and closer to X in the sense that for all $x \neq 0$, $F_n(x)$ is well approximated by $F(x)$. Alternatively, if we did not restrict convergence in distribution to the continuity points, we would have the strange case where a non-stochastic sequence $\{X_n\}$ converges to X under the non-stochastic definition of convergence but not converge in distribution.

We can extend each of these definitions to random vectors. For example, the sequence of random vectors $\{X_n\} \xrightarrow{as} X$ if each element of X_n converges almost surely to each element of X. The extension is analogous for convergence in probability. A sequence of random vectors converges into distribution to a random vector if we apply the definition above to the joint cumulative distribution function. Alternatively, the following theorem provides another characterization of multivariate convergence in distribution.

Theorem 3.1. *Cramér–Wold Device*

Let $\{Z_n\}$ be a sequence of k-dimensional random vectors. Then, $Z_n \xrightarrow{d} Z$ if and only if $\lambda' Z_n \xrightarrow{d} \lambda' Z$ for all $\lambda \in \mathbb{R}^k$.

How do these different definitions of stochastic convergence relate to each other? The next set of propositions lay out the relationships.

Proposition 3.1. *Convergence in mean-square implies convergence in mean*

Suppose $X_n \xrightarrow{ms} X$. Then, $X_n \xrightarrow{m} X$.

Proof. This follows from Jensen's inequality. Recall that if $h(\cdot)$ is a convex function, then

$$E[h(Y)] \geq h(E[Y]).$$

Set $h(z) = z^2$ and $Y = |X_n - X|$. It follows that

$$0 \leq E[|X_n - X|]^2 \leq E[|X_n - X|^2].$$

with $E[|X_n - X|^2] \to 0$. The result follows. □

Proposition 3.2. *Convergence in mean-square implies convergence in probability*

Suppose $X_n \xrightarrow{ms} X$. Then, $X_n \xrightarrow{p} X$.

Proof. This follows from Markov's inequality. Recall that for all $c > 0$,

$$P(Y \geq c) \leq E[Y]/c.$$

Fix $\epsilon > 0$. Set $c = \epsilon^2$ and $Y = |X_n - X|^2$. We have that

$$0 \leq P(|X_n - X|^2 \geq \epsilon^2) = P(|X_n - X| \geq \epsilon) \leq E[|X_n - X|^2]/\epsilon^2.$$

Taking limits of both sides, we get that

$$0 \leq \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) \leq \lim_{n\to\infty} E[|X_n - X|^2]/\epsilon^2.$$

and the result follows. □

Proposition 3.3. *Convergence in mean implies convergence in probability*

Suppose $X_n \xrightarrow{m} X$. Then, $X_n \xrightarrow{p} X$.

Proof. This follows from Markov's inequality analogously. □

Proposition 3.4. *Almost sure convergence implies convergence in probability*

Suppose $X_n \xrightarrow{as} X$. Then, $X_n \xrightarrow{p} X$.

Remark 3.5. *To prove convergence in probability, it is often easiest to prove convergence in mean-square. How do you show convergence in mean-square? Note that*

$$E[|X_n - X|^2] = V(X_n - X) + (E[X_n] - E[X])^2.$$

Therefore, $X_n \xrightarrow{ms} X$ if $V(X_n - X) \to 0$ and $E[X_n] - E[X] \to 0$.

Proposition 3.5. *Convergence in probability implies convergence in distribution*

Suppose $X_n \xrightarrow{p} X$. Then, $X_n \xrightarrow{d} X$.

Exercise 3.1. *Let $Y \sim N(0,1)$ and $Y_n = (-1)^n Y$. Does $Y_n \xrightarrow{d} Y$? Does $Y_n \xrightarrow{p} Y$?*

We conclude this section with two theorems that are very useful in deriving asymptotic distributions.

Theorem 3.2. *Slutsky's theorem*

Let c be a constant. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$. Then,

1. *$X_n + Y_n \xrightarrow{d} X + c$.*
2. *$X_n Y_n \xrightarrow{d} Xc$.*
3. *$X_n / Y_n \xrightarrow{d} X/c$ provided that $c \neq 0$.*

If $c = 0$, then $X_n Y_n \xrightarrow{p} 0$.

Theorem 3.3. *Continuous mapping theorem*

Let g be a continuous function. Then,

1. *If $X_n \xrightarrow{d} X$, then $g(X_n) \xrightarrow{d} g(X)$.*
2. *If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$.*

Proof. We provide the proof for (2) and the case where $X = a \in \mathbb{R}$. Let $\epsilon > 0$. Since $g(\cdot)$ is continuous at a, there exists some $\delta > 0$ such that

$$|x - a| < \delta \implies |g(x) - g(a)| < \epsilon.$$

The contrapositive of this is

$$|g(x) - g(a)| \geq \epsilon \implies |x - a| \geq \delta.$$

Substituting in X_n, it follows that

$$P(|g(X_n) - g(a)| \geq \epsilon) \leq P(|X_n - a| \geq \delta) \to 0.$$

The result follows. □

O_p and o_p Notation

Recall big-O and little-o notation for sequences of real numbers. Let $\{a_n\}$ and $\{g_n\}$ be sequences of real numbers. We have that

$$a_n = o(g_n) \quad \text{if} \quad \lim_{n\to\infty} \frac{a_n}{g_n} = 0$$

and

$$a_n = O(g_n) \quad \text{if} \quad \left|\frac{a_n}{g_n}\right| < M \quad \forall n.$$

Just like we extended the definition of non-stochastic convergence to sequences of random variable, we also extend big-O and little-o notation.

Definition 3.6. *Suppose $\{A_n\}$ is a sequence of random variables. We write*

$$A_n = o_p(G_n) \quad \textit{if} \quad \frac{A_n}{G_n} \xrightarrow{p} 0$$

and

$$A_n = O_p(G_n)$$

if for all $\epsilon > 0$, there exists $M \in \mathbb{R}$ such that $P(|\frac{A_n}{G_n}| < M) > 1 - \epsilon$ for all n.

Remark 3.6. *You'll often see someone write $X_n = X + o_p(1)$ to denote $X_n \xrightarrow{p} X$.*

Proposition 3.6. *If $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$.*

Proof. Since X is a random variable, there exists some $M > 0$ such that F_X is continuous at $-M, M$ and $P(|X| > M) = F_X(-M) + (1 - F_X(M)) < \epsilon/2$. Since $X_n \xrightarrow{d} X$, for all n large enough

$$|F_{X_n}(-M) - F_X(-M)| < \epsilon/4, \quad |F_{X_n}(M) - F_X(M)| < \epsilon/4.$$

And so, for all n large enough, we have that $P(|X_n| > M) < \epsilon$. □

Example 3.2. *Let $X_n \sim N(0, n)$. Then,*

$$X_n = O_p(n^{1/2}).$$

Why? We have that $X_n/n^{1/2} \sim N(0,1)$ for all n. For any $\epsilon > 0$, we can choose an M such that $P(|N(0,1)| < M) > 1 - \epsilon$. We also have that

$$X_n = o_p(n).$$

Why? We have that $X_n/n \sim N(0, 1/n)$. Note that

$$P(|N(0,1/n)| > \epsilon) = P(|N(0,1)| > n^{1/2}\epsilon) \to 0.$$

Alternatively, note that

$$E[(X_n/n - 0)^2] = V(X_n/n) = 1/n \to 0$$

and so, $X_n \xrightarrow{ms} 0$.

The Law of Large Numbers and Central Limit Theorem

There are two basic building blocks that we use to construct all asymptotic results. The first set of building blocks are laws of large numbers (LLNs). These show that sample averages converge to expectations under certain conditions. The second set of building blocks are central limit theorems (CLTs). These show that properly centered sample averages will converge in distribution to normal random variables. In this section, we provide several LLNs and CLTs that appear regularly.

Theorem 3.4. *Weak law of large numbers*

Let $X_1, \ldots, X_n$ be a sequence of random variables with $E[X_i] = \mu, V(X_i) = \sigma^2 < \infty$ for all i and $Cov(X_i, X_j) = 0$ for all $i \neq j$. Then,

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu.$$

Proof. By Chebyshev's inequality,

$$P(|\bar{X}_n - \mu| > \epsilon^2) \leq E[(\bar{X}_n - \mu)^2]/\epsilon^2 = \sigma^2/n\epsilon^2 \to 0.$$

Alternatively,

$$V(\bar{X}_n) = E[(\bar{X}_n - \mu)^2] = \sigma^2/n \to 0, \quad E[\bar{X}_n] = \mu$$

and so, $\bar{X}_n \xrightarrow{ms} \mu$ and the result follows. □

Theorem 3.5. *Chebyshev's weak law of large numbers*

Let $X_1, X_2, \ldots$ be a sequence of random variables with $E[X_i] = \mu_i, V(X_i) = \sigma_i^2$ and $Cov(X_i, X_j) = 0$ for all $i \neq j$. Define

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i, \quad \bar{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2$$

and assume that $\bar{\sigma}_n^2/n \to 0$. Then,

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0.$$

Proof. First, we have that

$$E[\bar{X}_n - \bar{\mu}_n] = 0.$$

Second, we have that

$$\begin{aligned} V(\bar{X}_n - \bar{\mu}_n) &= V(\bar{X}_n) \\ &= \frac{1}{n^2} \sum_{i,j} Cov(X_i, X_j) \\ &= \frac{1}{n^2} \sum_{i} \sigma_i^2 = \bar{\sigma}_n^2/n \to 0. \end{aligned}$$

Therefore, $\bar{X}_n - \bar{\mu}_n \xrightarrow{ms} 0$ and so, $\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0$. □

Theorem 3.6. *Strong law of large numbers*

If $X_1, X_2, \ldots$ are i.i.d. with $E[X_i] = \mu < \infty$, then

$$\bar{X}_n \xrightarrow{as} \mu.$$

Remark 3.7. *Note that for the weak law of large numbers, we only required the sequence of X_i's to be uncorrelated and also required finite second moments. For the strong law of large numbers, we required the X_i's to be i.i.d. but did not require any assumptions about second moments.*

Theorem 3.7. *Central limit theorem I*

Let $Y_1, Y_2, \ldots$ be an i.i.d. sequence of random variables with $E[Y_i] = 0, V(Y_i) = 1$ for all i. Then,

$$\sqrt{n}\bar{Y} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \xrightarrow{d} N(0,1).$$

Theorem 3.8. *Central limit theorem II*

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Then,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

This generalizes to random vectors. If $X_1, X_2, \ldots$ are i.i.d. random vectors with mean vector μ and covariance matrix Σ. Then,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \Sigma).$$

Exercise 3.2. *Let $W_i \sim \chi^2_{10}$ i.i.d. and define $\bar{W}_n = \frac{1}{n}\sum_{i=1}^n W_i$.*

1. *Show that $E[\bar{W}_n] = 10$.*
2. *Show that $\bar{W}_n \xrightarrow{p} 10$.*
3. *Show that $\frac{1}{n}\sum_{i=1}^n (W_i - \bar{W})^2 \xrightarrow{p} V(W_i)$.*
4. *Does $E[\frac{1}{n}\sum_{i=1}^n (W_i - \bar{W})^2] = V(W_i)$?*

Exercise 3.3. *Suppose $X_i \sim N_p(0, \Sigma)$ for $i = 1, \ldots, n$. Let $\alpha \in \mathbb{R}^p$ and define*

$$Y_n = \frac{\alpha' X_n' X_n \alpha}{\frac{1}{n-1}\sum_{i=1}^{n-1} \alpha' X_i' X_i \alpha}.$$

1. *Show that $Y_n \sim F_{1,n-1}$.*
2. *Show that $Y_n \xrightarrow{d} \chi^2_1$.*

The Delta Method

Suppose we have some estimator T_n of a parameter θ. We know that

$$T_n \xrightarrow{p} \theta$$

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

However, we are interested in estimating and conducting inference on $g(\theta)$, where g is some continuously differentiable function. A natural estimator is $g(T_n)$ and by the continuous mapping theorem, we know that

$$g(T_n) \xrightarrow{p} g(\theta).$$

Can we construct the asymptotic distribution of $g(T_n)$? That is,

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} ?$$

The **delta method** provides a general technique for constructing this asymptotic distribution.

Theorem 3.9. *The delta method*

Let Y_n be a sequence of random variables and let $X_n = \sqrt{n}(Y_n - a)$ for some constant a. Let $g(\cdot)$ be a continuously differentiable function. Suppose that

$$X_n = \sqrt{n}(Y_n - a) \xrightarrow{d} X \sim N(0, \sigma^2).$$

Then,

$$\sqrt{n}(g(Y_n) - g(a)) \xrightarrow{d} g'(a) N(0, \sigma^2).$$

Proof. By the mean value theorem,[4]

$$g(Y_n) = g(a) + (Y_n - a)g'(\tilde{Y}_n)$$

where $\tilde{Y}_n$ is some value between Y_n and a. Since $X_n \xrightarrow{d} X$, it follows that $Y_n \xrightarrow{p} a$. Since g is continuously differentiable, it follows that $g'(\tilde{Y}_n) \xrightarrow{p} g'(a)$ by the continuous mapping theorem. So, it follows that

$$\begin{aligned}\sqrt{n}(g(Y_n) - g(a)) &= g'(\tilde{Y}_n)\sqrt{n}(Y_n - a) \\ &= g'(\tilde{Y}_n)X_n \xrightarrow{d} g'(a)X\end{aligned}$$

by Slutsky's theorem. □

We can prove a similar result for random vectors. In the theorem above, replace everything with vectors. The result becomes

$$\sqrt{n}(g(Y_n) - g(a)) \xrightarrow{d} GN(0, \Sigma)$$

where

$$G = \frac{\partial g(a)}{\partial a'}.$$

Example 3.3. *Suppose X_i i.i.d. for $i = 1, \dots, n$ with mean 2 and variance 1. Then,*

$$\sqrt{n}(\bar{X} - 2) \xrightarrow{d} N(0, 1).$$

Let $g(z) = z^2$. The delta method tells us that

$$\sqrt{n}(\bar{X}^2 - 4) \xrightarrow{d} g'(2)N(0, 1) = N(0, 16).$$

Stationarity and Martingales

So far, each of the LLNs and CLTs we discussed relied on the random variables in the sequence to be uncorrelated or even independent. In this section, we briefly introduce a LLN-type result and CLT for dependent data. The presentation in this section closely follows Chapter 2 of Hayashi's *Econometrics*.[5]

A **stochastic process** is a sequence of random variables. A **time series** is a stochastic process whose indices are time measurements.

Definition 3.7. *A stochastic process is **strictly stationary** if the probability distribution of*

$$(X_t, X_{t+1}, \dots, X_{t+k})$$

is the same as the probability distribution of

$$(X_\tau, X_{\tau+1}, \dots, X_{\tau+k})$$

for all t, τ, k.

Definition 3.8. *A strictly stationary stochastic process is **ergodic** if for any two bounded functions $f : \mathbb{R}^k \to \mathbb{R}$ and $g : \mathbb{R}^k \to \mathbb{R}$, the following holds:*

$$\lim_{n\to\infty} E[f(X_t, \dots, X_{t+k})g(X_{t+n}, \dots, X_{t+n+k})] = E[f(X_t, \dots, X_{t+k})]E[g(X_t, \dots, X_{t+k})]$$

That is, sub-sequences separated by n time periods become independent as n grows large.

[4] Recall the mean value theorem? Let $g(\cdot)$ be a continuously differentiable function and WLOG, let $a < b$. There exists some $c \in (a, b)$ such that $g(b) = g(a) + g'(c)(b - a)$.

[5] This section is purely optional. Feel free to skip it.

Theorem 3.10. *Ergodic law of large numbers*

Suppose $\{X_i\}$ is a strictly stationary, ergodic stochastic process with $E[X_1] = \mu$. Then,

$$\bar{X}_n \xrightarrow{as} \mu.$$

Definition 3.9. *A stochastic process $\{Z_i\}$ is a **martingale** if*

$$E[Z_i | Z_{i-1}, \dots, Z_1] = Z_{i-1}$$

*for all $i \geq 2$. A stochastic process $\{Z_i\}$ with $E[Z_i] = 0$ for all i is a **martingale difference sequence** if*

$$E[Z_i | Z_{i-1}, \dots, Z_1] = 0$$

for all $i \geq 2$.

Theorem 3.11. *Ergodic stationary martingale difference CLT*

Let $\{X_i\}$ be a martingale difference sequence that is stationary and ergodic with

$$E[X_i X_i'] = \Sigma.$$

Then,

$$\sqrt{n}\bar{X}_n \xrightarrow{d} N(0, \Sigma).$$

References

DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*.

Hayashi, F. (2000). *Econometrics*.

Hogg, R., McKean, J. and A. Craig. (2012). *Introduction to Mathematical Statistics*.

van der Vaart, A. (2000). *Asymptotic Statistics*.

Chapter 4. Frequentist Inference

These notes provide a brief introduction to frequentist inference. It is intended to provide a simple, very high-level framework for thinking about many of the tools that will be discussed in 2120.

Contents

What is Frequentist Inference?

We begin by observing some data x_i for $i = 1, \ldots, n$ and assume that these data are the result of some random experiment. We model this random experiment with random variable X with support $\mathcal{X}$ and so, the data $\{x_i\}_{i=1}^n$ are realizations of X. We wish to use the data to learn something about the distribution of X, $F_X(x)$.

To do so, we construct a **statistical model**. A statistical model is a set of probability distributions indexed by a parameter set. That is, $\mathcal{F} = \{P_\theta(x) : x \in \mathcal{X}, \theta \in \Theta\}$ is a statistical model. A model is **parametric** if P can be indexed with a finite dimensional parameter set. Otherwise, it is **nonparametric**. The econometrician observes $\{x_i\}_{i=1}^n$ and wishes to make inferences about θ.

Frequentists assume that even though θ is unknown, we should view it as *fixed*. The data are modeled as random variables $X_1, \ldots, X_n$ drawn from the fixed, unknown distribution $F_\theta(x)$. Put in another way, frequentists model the random experiment as:

1. Nature draws realizations $x_1, \ldots, x_n$ from the distribution $F_\theta(x)$. These are the data.
2. The econometrician observes the data $x_1, \ldots, x_n$ and plugs them into her estimator, $\hat{\theta}(\cdot)$. Her estimate $\hat{\theta}^*$ is $\hat{\theta}(x_1, \ldots, x_n)$.

With this in mind, frequentists then perform the following thought experiment.

> Suppose I were to repeat the random experiment above many times. Each time I repeat the experiment, I obtain new data $x_1^b, \ldots, x_n^b$ and construct a new estimate using my estimator, $\hat{\theta}(x_1^b, \ldots, x_n^b) = \hat{\theta}^b$. What properties will the **sampling distribution** of my estimator have? That is, as $B \to \infty$, what properties will the distribution of $(\hat{\theta}^1, \ldots, \hat{\theta}^B)$ have?

For this reason, Bradley Efron and Trevor Hastie note that *"behaviorism"* would be a better name for frequentism because it better focuses attention on the emphasis placed the *behavior* of estimators in a **repeated random experiment**.

The philosophy of frequentist inference

Examples of experiments that try to estimate unknown parameters include:

- Michelson's experiment to estimate the speed of light
- In your high school physics class, dropping objects from different heights to estimate the acceleration due to gravity
- Sampling Americans randomly and asking them questions to estimate the latent popularity of a political candidate

Implicitly, frequentist methods are based on a type of experimentation more common in the natural sciences than in the social sciences. For example, frequentist methods are most natural when:

- It is always possible (though sometimes costly) to get more samples. Just make your RA work longer hours.
- It makes sense to talk about repeated independent experiments. A team at Harvard, a team at MIT, etc. each runs the same experiment separately, each writes a paper in *Science* reporting a point estimate $\hat{\theta}$, and a reader can compare all the experiments.
- The entire experimental procedure, including sample size, is specified before looking at any of the data.

It's common to apply frequentist methods to observational data, but some care is needed with the interpretation.[1] The possibility of running independent experiments is often more hypothetical than real.

Goals of frequentist inference

A frequentist statistician will observe the outcome of a random experiment and then report one or more of the following, depending on her research question:

1. A point estimate: her best guess of θ after looking at the data
2. A confidence set: a set of possible values that probably covers θ
3. The result of a hypothesis test: a judgment of whether a fact about θ is true

Frequentist Point Estimation

An **estimator**,[2,3] here denoted $\hat{\theta}$, is a function that maps the realized data into an estimate $\hat{\theta}^* = \hat{\theta}(x_1, \dots, x_n)$, which is a guess of the (fixed but unknown) vector θ. *Because the data are realizations of a random variable, the estimate is also a realization of a random variable.*

Classical frequentist point estimators include:

- **Classical method of moments.** The data are $X = (X_1, \dots, X_n)$. Assume that X_i are i.i.d. with density $f(x;\theta)$. Suppose we observe realizations $x = (x_1, \dots, x_n)$. We can derive population moments $E[X_i], E[X_i^2], \dots, E[X_i^k]$ as a function of θ. Meanwhile, in our sample, we observe sample moments $\frac{1}{n}\sum_i x_i, \frac{1}{n}\sum_i x_i^2, \dots, \frac{1}{n}\sum_i x_i^k$. By matching the first k sample moments to the first k population moments, we get a system of equations:

$$\int_{-\infty}^{\infty} x f(x;\theta)\, dx = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\vdots$$

$$\int_{-\infty}^{\infty} x^k f(x;\theta)\, dx = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

  The method of moments estimate $\hat{\theta}^*_{\text{MM}}$ is the value of θ that solves this system of equations. Typically, in order to prevent an overdetermined system, we choose k to be the dimension of θ.

[1] For example, what does it mean to get more independent samples from hospital admission data?

[2] An **estimator** is a process or function, while an **estimate** is the result of applying the estimator to a particular dataset.

[3] The estimate wears a hat. It's like how if a person is wearing a hat, you can't see him very well. —Wei Biao Wu

In order to use the classical method of moments, we need expressions for population moments as a function of θ. Strictly speaking, we don't need to assume anything more about the distribution of the X_i.

Example 4.1. *The data are $X = (X_1, \ldots, X_n)$. Assume that the X_i are i.i.d. with $X_i \sim \text{Poisson}(\lambda)$. Recall that the probability mass function is*

$$f(k;\lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Suppose we observe realizations $x = (x_1, \ldots, x_n)$. You can derive that $E[X] = \lambda$. Therefore, the classical method of moments estimate $\hat{\lambda}^*_{MM}$ is the sample mean, $\frac{1}{n}\sum_{i=1}^n x_i$.

- **Maximum likelihood (MLE).** Loosely speaking, rather than matching moments, this method matches the density function. Define the **likelihood** of the realized data given θ as the joint density or mass function evaluated at the realized data, $f(x_1, \ldots, x_n; \theta)$, and write it $L(\theta)$.[4] For discrete random variables, this is the probability mass function: the probability of observing the sample we saw. For continuous random variables, this is the probability density function.[5] The maximum likelihood estimate $\hat{\theta}^*_{\text{MLE}}$ is the value of θ that maximizes the likelihood, evaluated at the realized data:

$$\hat{\theta}^*_{\text{MLE}} = \arg\max_{\theta} L(\theta)$$

In common cases, we have tricks to help with this optimization problem. With i.i.d. data, it helps that we can rewrite $L(\theta)$ as a product of the likelihood for each observation:

$$L(\theta) = f(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

If the likelihood is strictly positive, $f(\cdot) > 0$, we can take logs.[6] The log of the likelihood function is called the **log likelihood** and usually denoted $\ell(\theta)$. In the i.i.d. case it simplifies as follows:

$$\ell(\theta) = \log L(\theta) = \log\left(\prod_{i=1}^n f(x_i; \theta)\right) = \sum_{i=1}^n \log f(x_i; \theta).$$

Then we can solve the optimization problem by maximizing the log likelihood, for example by taking first order conditions.[7] For many (but not all) common distributions, the second order conditions are guaranteed to hold.

Example 4.2. *It turns out that the maximum likelihood estimate of the Poisson parameter coincides with the method of moments estimate. We will now show this. As before, the data are $X = (X_1, \ldots, X_n)$, where the X_i are i.i.d. with $X_i \sim \text{Poisson}(\lambda)$, and we observe realizations $x = (x_1, \ldots, x_n)$. The likelihood function for each observation x_i is the probability mass function evaluated at x_i,*

$$f(x_i; \lambda) = P(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

Because of i.i.d. data, the log likelihood is

$$\ell(\lambda) = \sum_{i=1}^n \log\left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!}\right) = (\log \lambda)\left(\sum_{i=1}^n x_i\right) - \lambda n + \textit{constant}$$

[4] The dependence on the data is dropped for notational convenience.

[5] The probability of observing the exact realized sample is zero.

[6] In economics and statistics, 'log' means natural log, unless otherwise noted.

[7] For many distributions, closed-form solutions don't exist, and we need to use numerical methods.

where the 'constant' depends on data but not λ, and so is irrelevant to the optimization problem. Now solve by taking the first order condition.[8]

$$\frac{\partial \ell}{\partial \lambda} = \frac{1}{\lambda}\left(\sum_{i=1}^{n} x_i\right) - n = 0$$

which is solved at $\hat{\lambda}^*_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$, again the sample mean. (Now check the second order condition.)

Properties of frequentist point estimators

The random variable $\hat{\theta}(X_1, \ldots, X_n) - \theta$ has a mean, called the **bias**, and a covariance matrix (which some call the variance-covariance matrix). If it is one-dimensional, its standard deviation is called the **standard error**.[9] In simple parametric models, these are often easy to approximate, especially for asymptotically normal estimators.[10]

Frequentists generally look for the following properties in their estimators:

1. Consistency: As $n \to \infty$, $\hat{\theta}(X_1, \ldots, X_n) \xrightarrow{p} \theta$.
2. Unbiasedness: $E[\hat{\theta}(X_1, \ldots, X_n)] = \theta$.
3. Efficiency: $V(\hat{\theta}(X_1, \ldots, X_n))$ is minimized.[11]
4. $\sqrt{n}$-consistency and asymptotic normality:[12] There exists a constant matrix V such that, as $n \to \infty$, $\sqrt{n}(\hat{\theta}(X_1, \ldots, X_n) - \theta) \xrightarrow{d} N(0, V)$.

There are other properties that may be desirable in practice. For example, econometricians often trade off some amount of efficiency to make their estimators more robust to model misspecification.[13]

Exercise 4.1. *The data are $X = (X_1, \ldots, X_n)$. Assume that the X_i are i.i.d. with $X_i \sim U[0, c]$, where c is unknown. Suppose we observe realizations $x = (x_1, \ldots, x_n)$.*

1. *Derive the method of moments estimate of c.*
2. *Is the method of moments estimate unbiased? Consistent? Asymptotically normal?*[14]
3. *Derive the maximum likelihood estimate of c. (Careful!)*
4. *Derive the cdf of the maximum likelihood estimate.*
5. *Is the maximum likelihood estimate unbiased? Consistent? Asymptotically normal?*

Confidence Sets

Definition 4.1. *The data are $X = (X_1, \ldots, X_n)$. A $1 - \alpha$ **confidence set** $\mathcal{C}(\cdot) \subset \Theta$ is a set-valued function of data with the property*

$$P(\theta \in \mathcal{C}(X)) = 1 - \alpha.$$

[8] What if all the x_i are zero?

[9] In multiple dimensions, the standard error of a component is the standard deviation along that dimension.

[10] In more complex models, approximations may rely on techniques like the bootstrap or the jackknife.

[11] This is defined loosely here; for more detail, look up minimum-variance unbiased estimators and the Cramér-Rao bound in your favorite statistics textbook.

[12] This isn't prized for its own sake, but it makes common techniques easy to apply.

[13] More on this when we discuss quasi-MLE and the generalized method of moments in 2120 and 2140.

[14] For a refresher on the tools used to prove consistency and asymptotic normality, see the note on asymptotics.

In one dimension, this is usually a **confidence interval**. The interpretation of this object is subtle. Each independent experiment delivers a different dataset, and each dataset generates a confidence set. When you pick up *Science* and read about the independent identical experiments that were conducted, expect the proportion of experiments whose confidence set includes θ to be $1 - \alpha$.

Even taking a model as given, there are many ways to construct a confidence set. We prefer small confidence sets (by Lebesgue measure, for example) and confidence sets centered at the point estimate.

Example 4.3. *The data are $X = (X_1, \dots, X_n)$. Assume that the X_i are i.i.d. with $X_i \sim N(\mu, \sigma^2)$, where μ is unknown and σ^2 is known, and write their mean as $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. To form a confidence interval for μ, note that*

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1).$$

*This object is called a **pivot**: it is a function of the parameters and data, and its distribution is known. Denote the cdf of the standard normal distribution as Φ. It follows that*

$$P\left(-\Phi^{-1}(1-\alpha/2) \leq \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \leq \Phi^{-1}(1-\alpha/2)\right) = 1 - \alpha$$

which, after some manipulation, yields

$$P\left(\mu \in \left[\bar{X}_n - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\sigma^2}{n}}, \bar{X}_n + \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\sigma^2}{n}}\right]\right) = 1 - \alpha.$$

Denote the realized data by $x_1, \dots, x_n$. Using our data, we report a $1 - \alpha$ confidence interval[15] *of*

$$\left[\frac{1}{n}\sum_{i=1}^n x_i - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\sigma^2}{n}}, \frac{1}{n}\sum_{i=1}^n x_i + \Phi^{-1}(1-\alpha/2)\sqrt{\frac{\sigma^2}{n}}\right].$$

This notion can be extended to an **asymptotic confidence set**, which satisfies $P(\theta \in \mathcal{C}(X)) \approx 1 - \alpha$ when n is large.[16]

Contrast this with the Bayesian notion of a credible set, which is (essentially) a parsimonious description of the posterior.

Hypothesis Testing

The goal of hypothesis testing is to use the data to answer a yes-no question; this is easier said than done. The machinery of hypothesis testing is built on a **null hypothesis** $H_0 : \theta \in \Theta_0$ and an **alternative hypothesis** $H_A : \theta \in \Theta_A$, where Θ_0 and Θ_A are disjoint subsets of Θ. Think of the statistician as a judge: after observing the outcome of a random experiment, she must adjudicate between the null and alternative hypotheses.[17]

Hypothesis tests are defined by a **decision rule** d mapping the realized data into $\{0, 1\}$, where 0 means "accept H_0", and 1 means "reject H_0 in favor of H_A". Many common hypothesis tests define a **test statistic** $T(x)$ to

[15] Letting $\alpha = 0.05$, as is common, we get that $\Phi^{-1}(1-\alpha/2) \approx 1.96$, a magic number in empirical work.

[16] Because of the central limit theorem, the mechanics of Example 4.3 often carry over.

[17] As with the presumption of innocence in criminal cases, the judge is conventionally obligated to privilege the null hypothesis over the alternative hypothesis.

map the realized data x into $\mathbb{R}$, and a fixed **rejection region** R, and use the following decision rule:

$$d(x) = \begin{cases} 0, & T(x) \notin R \\ 1, & T(x) \in R \end{cases}.$$

We evaluate hypothesis tests by the two ways they can go wrong:

- Type I error (false rejection): if we reject the null hypothesis but the null hypothesis is in fact true ($\theta \in \Theta_0$). The probability[18] of making a Type I error is known as **significance** and usually denoted α.
- Type II error (false acceptance): if we accept the null hypothesis but in fact the alternative hypothesis is true ($\theta \in \Theta_A$). The probability of avoiding a Type II error is known as **power** and usually denoted $1 - \beta$.

How do we determine the rejection region R? The frequentist statistician starts with a significance level, such as $\alpha = 0.05$, and then chooses a rejection region $R(\alpha)$ such that the significance of the test is α. Given the data, the **p-value** is the smallest significance level at which the null hypothesis would be rejected:

$$p(x) = \inf\{\alpha : T(x) \in R(\alpha)\}$$

Generalized likelihood ratio test

The generalized likelihood ratio test is one common hypothesis test. Recall the **likelihood** from maximum likelihood estimation, $L(\theta)$, which depends on the data. In the generalized likelihood ratio test, the test statistic is the ratio of likelihoods:

$$\lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta' \in \Theta_0 \cup \Theta_A} L(\theta')}.$$

and the rule is to reject if $\lambda \leq c$, where c is a constant.[19] To determine c given a significance level, note that under regularity conditions, if the null hypothesis is true and the sample is large, then $-2 \log \lambda$ is approximately chi-squared with $\dim(\Theta_0 \cup \Theta_A) - \dim \Theta_0$ degrees of freedom. You can work out the rest.

Duality of confidence sets and hypothesis tests

The results below consider tests of a simple null hypothesis, where $\Theta_0 = \{\theta_0\}$.

Theorem 4.1. *Suppose that for each $\theta_0 \in \Theta$, there exists a test of the simple null hypothesis $H_0 : \theta = \theta_0$ with significance α. Denote its decision rule by d_{θ_0}. Then*

$$\mathcal{C}(x) = \{\theta_0 : d_{\theta_0}(x) = 0\}$$

is a $1 - \alpha$ confidence set for θ.

That is, we can start with a hypothesis test and construct a confidence set by asking what simple null hypotheses would be accepted. The following result lets us go the other way:

Theorem 4.2. *Let $\mathcal{C}(x)$ be a $1 - \alpha$ confidence set for θ. Then the following test is a hypothesis test of the null hypothesis $H_0 : \theta = \theta_0$ with significance α:*

$$\textit{accept } H_0 \textit{ if } \theta_0 \in \mathcal{C}(x), \textit{ reject } H_0 \textit{ otherwise.}$$

[18] To keep our frequentist credentials, note that 'probability' here means the long-run rate across many independent experiments.

[19] Here, λ is the test statistic, and $(-\infty, c]$ is the rejection region.

References

Newey, W., and D. McFadden. (1994). "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics*. `https://doi.org/10.1016/S1573-4412(05)80005-4`.

Rice, J. (2007). *Mathematical Statistics and Data Analysis*, 3rd ed.[20]

Wasserman, L. (2004). *All of Statistics*. `https://doi.org/10.1007/978-0-387-21736-9`.

[20] I only used this book because my undergraduate class used it. Any undergraduate textbook in mathematical statistics will do.

Chapter 5. Bayesian Inference

These notes provide a brief introduction to Bayesian inference. It is intended to provide a simple, very high-level framework for thinking about many of the tools that will be discussed in 2120. [1]

Contents

[1] These notes draw heavily on Chapters 2-3 of *Computer Age Statistical Inference* by Bradley Efron and Trevor Hastie and Gary Chamberlain's lecture note 5 for Ec 2120 for these notes. This is an excellent textbook, and the pdf is freely available online!

What is Bayesian Inference?

We begin by observing some data x_i for $i = 1, \dots, n$ and assume that these data are the result of some random experiment. We model this random experiment with random variable X with support $\mathcal{X}$ and so, the data $\{x_i\}_{i=1}^n$ are realizations of X. We wish to use the data to learn something about the distribution of X, $F_X(x)$.

To do so, we construct a **statistical model**. A statistical model is a set of probability distributions indexed by a parameter set. That is, $\mathcal{F} = \{P_\theta(x) : x \in \mathcal{X}, \theta \in \Theta\}$ is a statistical model. A model is **parametric** if P can be indexed with a finite dimensional parameter set. Otherwise, it is **nonparametric**. The econometrician observes $\{x_i\}_{i=1}^n$ and wishes to make inferences about θ.

Example 5.1. *Suppose our statistical model is the set of normal distributions with variance equal to one. Then,* $\mathcal{X} = \mathbb{R}$, $\Theta = \mathbb{R}$ *and*

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}.$$

Example 5.2. *Suppose our statistical model is the set of Poisson distributions. Then,* $\mathcal{X} = \mathcal{N}$, $\Theta = \mathbb{R}_+$ *and*

$$f_\theta(x) = e^{-\theta}\theta^x / x!.$$

So, both frequentists and bayesians begin with a probability model and wish to learn about the parameter θ. What makes them different? Go back to the definition of a statistical model. Suppose we have a "good" statistical model. That is, $F_X(x) \in \mathcal{F}$ and so, there exists some $\theta^* \in \Theta$ such that $F_X(x) = F_{\theta^*}(x)$. *The whole point of statistical inference is that* θ^* *is unknown.* I think the key difference between frequentists and bayesians is in how they model an unknown θ^* and what that, in turn, implies for how inference should be conducted.[2]

Frequentists assume that even though θ^* is unknown, we should view it as *fixed*. The data are modeled as random variables $X_1, \dots, X_n$ drawn from the fixed, unknown distribution $F_{\theta^*}(x)$. Put in another way, frequentists model the random experiment as:

1. Nature draws realizations $x_1, \dots, x_n$ from the distribution $F_{\theta^*}(x)$. These are the data.
2. The econometrician observes the data $x_1, \dots, x_n$ and plugs them into her estimator, $\hat{\theta}(\cdot)$. Her estimate of $\hat{\theta}^*$ is $\hat{\theta}(x_1, \dots, x_n)$.

With this in mind, frequentists then perform the following thought experiment.

> Suppose I were to repeat the random experiment above many times. Each time I repeat the experiment, I obtain new data $x_1^b, \dots, x_n^b$ and construct a new estimate using my estimator, $\hat{\theta}(x_1^b, \dots, x_n^b) = \hat{\theta}^b$. What properties will the **sampling distribution** of my estimator have? That is, as $B \to \infty$, what properties will the distribution of $(\hat{\theta}^1, \dots, \hat{\theta}^B)$ have?

For this reason, Bradley Efron and Trevor Hastie note that *"behaviorism"* would be a better name for frequentism because it better focuses attention

[2] If you ask some people, they will emphasize that the difference lies in how frequentists and bayesians interpret probability.

on the emphasis placed the *behavior* of estimators in a **repeated random experiment**.[3] An example of a desirable property for frequentists is **unbiasedness**. An estimator $\hat{\theta}(\cdot)$ is **unbiased** if $E[\hat{\theta}(X_1, \ldots, X_n)] = \theta^*$. Note that this expectation is taken over the sampling distribution of the estimator $\hat{\theta}$.

Bayesians, on the other hand, prefer to model the unknown $\hat{\theta}^*$ as a random variable itself. $\hat{\theta}^*$ is a random variable that has its own distribution, $\Pi(\theta)$. This is called the **prior distribution**. The random experiment then has an extra step:

1. Nature draws θ^* from the prior distribution, $\Pi(\theta)$. This is unobserved.
2. Nature draws realizations $x_1, \ldots, x_n$ from the distribution $F_{\theta^*}(x)$. These are the data.
3. The econometrician observes $x_1, \ldots, x_n$ and plugs them into her estimator, $\hat{\theta}(\cdot)$. Her estimate is $\hat{\theta}(x_1, \ldots, x_n)$.

Clearly, the prior distribution will be an important part of bayesian inference. How should we think about it? For now, think of the prior distribution as encoding *prior information* about the parameter θ available to the econometrician prior to observing the data. This may come from prior experiments, observational studies, or economic theory.

What is the point of adding this additional layer? The payoff comes from the use of **Bayes' rule**. Bayes' rule provides a logically consistent rule for combining prior information with the observed data. Let $x = (x_1, \ldots, x_n)$ and let $f_\theta(x)$ denote the density associated with the distribution $F_\theta(x)$ and $\pi(\theta)$ is defined analogously. Bayes' rule tells us

$$\pi(\theta|x) = \frac{f_\theta(x)\pi(\theta)}{f(x)}$$

where $f(x) = \int_\Theta f_\theta(x)\pi(\theta)\, d\theta$ is the **marginal density** of X. $f_\theta(x)$ is the **likelihood function**. We call $\pi(\theta|x)$ the **posterior density** of θ and it is the key object of bayesian inference.[4] The bayesian then uses the posterior distribution to make inferences about θ. For example, a common object of interest is the "posterior expectation of θ given the data x"

$$E[\theta|x].$$

However with the posterior distribution, the bayesian immediately answers all possible questions about θ. She could compute $E[\theta|x], Med(\theta|X), P(\theta < \tilde{\theta}|X)$ and so on. It is critical to note that in the posterior density, x is *fixed* at its realized value and θ varies over Θ. In this sense, bayesian inference is completely *conditional on the observed data*.

Conjugate Priors

As mentioned above, the choice of the prior distribution is the key step of bayesian inference. Once we have a prior distribution and a likelihood function, the only computational step is to use Bayes' rule to form the posterior. While it sounds simple, this can often be a mess unless we carefully choose the prior distribution for a given likelihood function.

As a result, an important tool in bayesian inference are **conjugate priors**. A prior distribution is **conjugate** for a given likelihood function if the associated posterior distribution is in the same family of distributions as the prior.

[3] Another way of thinking about this is: In frequentist calculations, θ^* is fixed and the data varies over different possible realizations conditional on θ^*.

[4] You will often see Bayes' rule written as

$$\pi(\theta|x) \propto f_\theta(x)\pi(\theta)$$

where $\propto$ means "is proportional to." In English Bayes' rule says, "the posterior is proportional to the likelihood times the prior."

The rest of this section covers some common conjugate priors that you will encounter throughout the first year econometrics sequence and in other areas of economics.

Normal-Normal model

The data are $X = (X_1, \dots, X_n)$. We assume that conditional on θ, the X_i are i.i.d. with

$$X_i \sim N(\mu, \sigma^2).$$

σ^2 is fixed and assumed known. It is useful to define the **precision** as $\lambda_\sigma = 1/\sigma^2$. The parameter space is $\Theta = \mathbb{R}$. Suppose we observe realizations $x = (x_1, \dots, x_n)$. The likelihood function is then

$$\begin{aligned} f_\mu(x) &= f(x|\mu) \\ &= \prod_{i=1}^n f(x_i|\mu) \\ &\propto \prod_{i=1}^n \exp(-\frac{1}{2}\lambda_\sigma(x_i-\mu)^2) \\ &\propto \exp(-\frac{1}{2}\lambda_\sigma \sum_{i=1}^n (x_i-\mu)^2). \end{aligned}$$

The prior distribution for μ is also normal. We assume that

$$\mu \sim N(m, \tau^2).$$

Again, it is useful to define the **prior precision** as $\lambda_\tau = 1/\tau^2$. We have that

$$\pi(\mu) \propto \exp(-\frac{1}{2}\lambda_\tau(\mu-m)^2).$$

The posterior distribution is given by Bayes' rule. We have that[5]

$$\begin{aligned} \pi(\mu|x) &\propto f_\mu(x)\pi(\mu) \\ &\propto \exp(-\frac{1}{2}\lambda_\sigma \sum_{i=1}^n (x_i-\mu)^2)\exp(-\frac{1}{2}\lambda_\tau(\mu-m)^2) \\ &\propto \exp\left(-\frac{\lambda_\sigma}{2}\sum_{i=1}^n (x_i^2 - 2x_i\mu + \mu^2) - \frac{\lambda_\tau}{2}(\mu^2 - 2\mu m + m^2)\right) \\ &\propto \exp\left(-\frac{n\lambda_\sigma + \lambda_\tau}{2}\mu^2 + 2\frac{\lambda_\sigma \sum_{i=1}^n x_i + \lambda_\tau m}{2}\mu\right) \\ &\propto \exp\left(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu^2 - 2\frac{\lambda_\sigma \sum_{i=1}^n x_i + \lambda_\tau m}{n\lambda_\sigma + \sigma_\tau}\mu)\right) \\ &\propto \exp\left(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu^2 - 2\frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}\mu)\right) \\ &\propto \exp\left(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu^2 - 2\frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}\mu + (\frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau})^2)\right) \\ &\propto \exp\left(-\frac{n\lambda_\sigma + \lambda_\tau}{2}(\mu - \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau})^2\right). \end{aligned}$$

And so, we have shown that the posterior distribution is also normally distributed with posterior mean

$$E[\mu|x] = \frac{n\lambda_\sigma \bar{x} + \lambda_\tau m}{n\lambda_\sigma + \lambda_\tau}$$

[5] The key to making this calculation easy is to remember that the posterior density is a function of μ. Since x, m are constants, we can drop them along the way.

and posterior precision

$$\bar{\lambda}_\tau = n\lambda_\sigma + \lambda_\tau.$$

What is the interpretation of the posterior mean? It is a weighted average of the sample mean and the prior mean in which the weights are the precisions. Therefore, if λ_τ is large and the prior has a low variance, the prior mean receives a larger weight. Alternatively, we can interpret this as "shrinking" the posterior mean towards the prior.[6]

We could have derived this using our results from the multivariate normal distribution. We have that

$$X|\mu \sim N(\mu, \sigma^2 I_n).$$

You can show that the marginal distribution of X is given

$$X \sim N(m, (\sigma^2 + \tau^2)I_n)$$

and that the joint distribution of X, μ is given by

$$\begin{pmatrix} X \\ \mu \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ m \end{pmatrix}, \begin{pmatrix} (\sigma^2+\tau^2)I_n & \tau^2 l \\ \tau^2 l' & \tau^2 \end{pmatrix}\right)$$

where l is a $n \times 1$ vector of ones. It then follows that

$$\mu|X = x \sim N(m + \frac{\tau^2}{\sigma^2+\tau^2} l' I_n (x-m), \tau^2 - \tau^2(\sigma^2+\tau^2)^{-1}\tau^2 l' l).$$

This is the same as the result we derived.

Exercise 5.1. *Use the properties of the multivariate normal distribution to derive the posterior distribution.*

Beta-Bernoulli model

The data are $X = (X_1, \ldots, X_n)$. We assume that conditional on θ, the X_i are i.i.d. with

$$P(X_i = 1|\theta) = \theta, \quad P(X_i = 0|\theta) = 1 - \theta.$$

The parameter space is $\Theta = [0, 1]$. Suppose we observe realizations $x = (x_1, \ldots, x_n)$. The likelihood function is then

$$\begin{aligned} f_\theta(x) &= f(x|\theta) \\ &= P(X = x|\theta) \\ &= \prod_{i=1}^n P(X_i = x_i|\theta) \\ &= \prod_{i=1}^n \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{n_1}(1-\theta)^{n_0} \end{aligned}$$

where $n_1 = \sum_{i=1}^n y_i$ and $n_0 = \sum_{i=1}^n (1 - y_i) = n - n_1$. The prior distribution is a **beta distribution** with parameters $a, b > 0$. Its support is over $[0, 1]$ with density

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.$$

[6] If you are familiar with machine learning jargon, you can write down a bayesian model similar to this to motivate Ridge regression.

The prior mean and variance are

$$E[\theta] = \frac{a}{a+b}, \quad V(\theta) = \frac{a}{a+b}\frac{b}{a+b}\frac{1}{a+b+1}.$$

The posterior distribution is given by Bayes' rule. We have that

$$\begin{aligned}\pi(\theta|x) &\propto f_\theta(x)\pi(\theta)\\ &\propto \theta^{a+n_1-1}(1-\theta)^{b+n_0-1}.\end{aligned}$$

Therefore, the posterior distribution is also a beta distribution and has parameters $a+n_1, b+n_0$. The posterior mean is then

$$E[\theta|x] = \frac{a+n_1}{a+b+n} = \lambda\frac{n_1}{n} + (1-\lambda)\frac{a}{a+b}$$

where $\lambda = \frac{n}{a+b+n}$. In other words, the posterior mean is a convex combination of the sample mean n_1/n and the prior mean $a/(a+b)$. Note that if $a+b$ is small relative to n, then most of the weight is placed on the sample mean.

What happens as $a, b \to 0$? The prior becomes

$$\pi(\theta) \propto \theta^{-1}(1-\theta)^{-1}.$$

This is not a probability density as it integrates to ∞ over $[0,1]$. We call this an **improper prior**. However, the associated posterior distribution is well-defined. In this case, the posterior distribution is again a beta distribution but with parameters, n_1, n_0. For this improper prior,

$$E[\theta|x] = \frac{n_1}{n} = \bar{x}$$

That is, the posterior conditional expectation coincides with the sample average (i.e., the frequentist estimate of θ).

Multinomial-Dirichlet model

The data are $X = (X_1, \ldots, X_n)$. Each X_i takes on a discrete set of values $\{\alpha_j : j = 1, \ldots, J\}$. We assume that conditional on θ, the X_i are i.i.d. with

$$P(X_i = \alpha_j|\theta) = \theta_j \quad \text{for } j = 1, \ldots, J.$$

The parameter space is the unit simplex on $\mathbb{R}^J$ with

$$\Theta = \{\theta \in \mathbb{R}^J : \theta_j \geq 0, \sum_{j=1}^{J}\theta_j = 1\}.$$

We observe realizations $x = (x_1, \ldots, x_n)$. The likelihood function is

$$\begin{aligned}f_\theta(x) &= f(x|\theta)\\ &= \prod_{i=1}^{n} P(X_i = x_i|\theta)\\ &= \prod_{i=1}^{n}\prod_{j=1}^{J}\theta_j^{1(x_i=\alpha_j)}\\ &= \prod_{j=1}^{J}\theta_j^{n_j}\end{aligned}$$

where $n_j = \sum_{i=1}^{n} 1(x_i = \alpha_j)$ for $j = 1, \ldots, J$.

The prior distribution is a **Dirichlet distribution** with parameters $a_1, \ldots, a_J > 0$. The Dirichlet distribution is a generalization of the beta distribution. Its support is over the unit simplex in $\mathbb{R}^J$ and has density

$$\pi(u_1, \ldots, u_J) \propto \prod_{j=1}^{J} u_j^{a_j - 1}.$$

The posterior distribution is then given by Bayes' rule. We have that

$$\begin{aligned} \pi(\theta|x) &\propto f_\theta(x)\pi(\theta) \\ &\propto \prod_{j=1}^{J} \theta_j^{a_j + n_j - 1}. \end{aligned}$$

The posterior distribution is also Dirichlet but with parameters $a_j + n_j$ for $j = 1, \ldots, J$. As in the Beta-Bernoulli model, we can consider the improper prior with $a_j \to 0$ for each $j = 1, \ldots, J$. With this improper prior, the posterior distribution remains Dirichlet and has parameters $n_1, \ldots, n_J$.

It turns out that we can represent the Dirichlet distribution using independent gamma distributed random variables. This is very useful in deriving several properties of the Dirichlet distribution and in simulations. The **gamma distribution** with **shape parameter** $a > 0$ and **scale parameter** $b > 0$ has density

$$g(u) \propto u^{a-1} \exp(-u/b)$$

with support over $u > 0$. The gamma distribution has the useful property that if Q_j are independent gamma distributed with parameters (a_j, b), then their sum $\sum_j Q_j \sim gamma(\sum_j a_j, b)$.

Suppose $Q_j \sim gamma(a_j, 1)$ for $j = 1, \ldots, J$ and $Q_1, \ldots, Q_j$ are independent. Let $S = \sum_{j=1}^{J} Q_j$. Define

$$R = (Q_1/S, \ldots, Q_J/S)$$

and one can show that $R \sim Dirichlet(a_1, \ldots, a_J)$. For the case $J = 2$, we have that

$$R = (Q_1/(Q_1 + Q_2), Q_2/(Q_1 + Q_2))$$

where $Q_1/(Q_1 + Q_2) \sim beta(a_1, a_2)$.

For the posterior distribution of θ, we can represent it as

$$\theta|x \sim \left(\frac{Q_1}{\sum_{j=1}^{J} Q_j}, \ldots, \frac{Q_J}{\sum_{j=1}^{J} Q_j} \right)$$

where each Q_j are mutually independent gamma random variables with parameters $a = n_j + a_j - 1, b = 1$. So a component θ_j can be represented as

$$\theta_j|x \sim \frac{Q_j}{Q_j + \sum_{k \neq j} Q_k}$$

and so, $\theta_j \sim beta(n_j + a_j, \sum_{k \neq j} n_k + a_k)$.

Exchangeability and de Finetti's Theorem

So far, we have assumed that there is some prior distribution π over θ and that conditional on θ, the observed data are i.i.d. de Finetti's Theorem, also known as the Representation Theorem, justifies this setup. de Finetti's Theorem and related generalizations show that if a sequence of random variables $X_1, \ldots, X_n$ are **exchangeable**, then there exists a parameter θ and a prior distribution π for θ such that the elements of the sequence are i.i.d. conditional on θ. This is a powerful result.

Definition 5.1. *A finite sequence of random variables $X_1, \ldots, X_n$ is **exchangeable** if its joint distribution $F(\cdot)$ satisfies*

$$F(x_1, \ldots, x_n) = F(x_{p(1)}, \ldots, x_{p(n)})$$

*for all realizations $(x_1, \ldots, x_n)$ and all permutations p of $\{1, \ldots, n\}$. Any infinite sequence of random variables is **exchangeable** if every finite subsequence is exchangeable.*

Remark 5.1. *Note that exchangeability is a weaker condition than i.i.d. If $X_1, \ldots, X_n$ are i.i.d., then the sequence is exchangeable. However, the elements of an exchangeable sequence are identically distributed but need not be independent.*

Example 5.3. *Polya's Urn*

Consider an urn with b black balls and w white balls. Draw a ball and note its color. Replace the ball in the urn and add a additional balls of the same color to the urn. Let $X_i = 1$ if the i-th drawn ball is black and $X_i = 0$ if it is white. The sequence $X_1, X_2, \ldots$ is exchangeable. For example,

$$\begin{aligned} f(1,1,0,1) &= \frac{b}{b+w}\frac{b+a}{b+w+a}\frac{w}{b+w+2a}\frac{b+2a}{b+w+3a} \\ &= \frac{b}{b+w}\frac{w}{b+w+a}\frac{b+a}{b+w+2a}\frac{b+2a}{b+w+3a} \\ &= f(1,0,1,1). \end{aligned}$$

Theorem 5.1. *de Finetti's Theorem*

Let $X_1, X_2, \ldots$ be an exchangeable sequence. Then, there exists a random variable Θ with cdf $F_\Theta(\cdot)$ such that

$$f(x_1, \ldots, x_n) = \int_0^1 \theta^{n_1}(1-\theta)^{n-n_1}\, dF_\Theta(\theta)$$

where

$$n_1 = \sum_{i=1}^n x_i$$

and

$$\Theta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n X_i$$

with $F_\Theta(\theta) = \lim_{n\to\infty} P(\frac{1}{n}\sum_{i=1}^n X_i \leq \theta)$.

It is as if the sequence of Bernoulli random variables are i.i.d. conditional on Θ. Moreover, the distribution of Θ is determined by the limiting distribution of the sample frequency. We can view F_Θ as a prior distribution. How do we interpret this? It provides us with a way to think about the prior distribution. By de Finetti's Theorem, the prior distribution F_Θ is determined by the limiting distribution of the sample frequency and so, we can view it as reflecting the researcher's subjective beliefs about the long-run frequency. de Finetti's Theorem generalizes in many ways. See, for instance, Diaconis (1988) for more results.

References

Diaconis, P. (1988). Recent Progress on de Finetti's Notions of Exchangeability. *Bayesian Statistics*.

Efron, Bradley and Trevor Hastie. (2016). *Computer Age Statistical Inference*.

Gelman et al. (2013). *Bayesian Data Analysis*.

Geweke, John. (2005). *Contemporary Bayesian Econometrics and Statistics*.

Lauritzen, Steffen. Exchangeability and de Finetti's Theorem. Lecture notes 2007.

Poirier, Dale. (2011). Exchangeability, Representation Theorem and Subjectivity. *The Oxford Handbook of Bayesian Econometrics*.