

Econ 623 Econometrics II

Topic 1: Review of Linear Algebra, Statistics, Asymptotic Theory, Computation, Measure Theory

1 Linear Algebra

- A matrix is a rectangular array: $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ & & \vdots & \\ a_{n1} & a_{n1} & \cdots & a_{nK} \end{bmatrix}$. A is an $n \times K$ matrix.
- If $n = K$, A is a square matrix. If $a_{ij} = a_{ji}$, A is a symmetric matrix. If $a_{ij} = 0 \forall i \neq j$, A is a diagonal matrix. If, in addition, $a_{ii} = 1 \forall i$, A is an identity matrix (often written as I). If $a_{ij} = 0 \forall i > j$, A is a triangular matrix.
- If $K = 1$, A is a column vector. If $n = 1$, A is a row vector.
- $B = A'$ (transpose of A) if $b_{ij} = a_{ji}, \forall i, j$. So $(A')' = A$. The transpose of a symmetric matrix is itself. The transpose of a column vector is a row vector.
- $C = A \pm B$ if $c_{ij} = a_{ij} \pm b_{ij}$. Obviously A and B are of the same dimension.
- $(A + B) + C = A + (B + C)$. $(A + B)' = A' + B'$.

- Let a and b be two column vectors. The inner product of a and b is $a'b = \sum_{i=1}^n a_i b_i$, a scalar. The outer product of a and b is ab' , which is a matrix with the ij th element being $a_i b_j$.
- To multiply two matrices, the number of columns in the first matrix must be same as the number of rows in the second matrix.
- It is easy to see that $AI = IA = A$. $(AB)C = A(BC)$. $A(B + C) = AB + AC$. $(AB)' = B'A'$. $A0 = 0$. But $AB \neq BA$.
- If λ is a scalar, B is a matrix, then $\lambda B = (\lambda \times b_{ij})$. Similar, we can define the product of a scalar and a vector.
- Suppose a_1, \dots, a_K are all column vectors, and $\lambda_1, \dots, \lambda_K$ are all scalars, then $c = \sum_{i=1}^K \lambda_i a_i$ is a linear combination of the column vector.
- Let ι be a column vector of 1s. Let a be a column vector. Then $\sum_{i=1}^n a_i = \iota'a$ and $\bar{a} = \frac{1}{n} \sum_{i=1}^n a_i = \frac{1}{n} \iota'a$. Also $\sum_{i=1}^n a_i^2 = a'a$.
- Let X be an $n \times K$ matrix whose i th column is x_i . Then $[X'X]_{ij} = x_i'x_j$ and $X'X = \sum_{i=1}^n x_i x_i'$.

- A symmetric matrix M is **idempotent** iff $M^2 \equiv MM = M$.
- An example: $M^0 = I - \frac{1}{n} \mathbf{1}\mathbf{1}'$. Understanding M^0 . Suppose x is a column vector.

$$\bar{x}\mathbf{1} = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix} = \bar{x} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \bar{x}\mathbf{1} = \frac{1}{n} \mathbf{1}'x = \frac{1}{n} \mathbf{1}'x.$$

$$\begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix} = x - \bar{x}\mathbf{1} = x - \frac{1}{n} \mathbf{1}'x = \left[I - \frac{1}{n} \mathbf{1}\mathbf{1}' \right] x = M^0 x.$$

So M^0 transform data to deviations from their mean. Furthermore,

$$\sum_{i=1}^n (x_i - \bar{x})^2 = (x - \bar{x}\mathbf{1})'(x - \bar{x}\mathbf{1}) = (M^0 x)'(M^0 x) = x' M^0 M^0 x = x' M^0 x$$

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = x' M^0 y.$$

$$\begin{bmatrix} \sum_{i=1}^n (x_i - \bar{x})^2 & \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}) & \sum_{i=1}^n (y_i - \bar{y})^2 \end{bmatrix} = \begin{bmatrix} x' M^0 x & x' M^0 y \\ y' M^0 x & y' M^0 y \end{bmatrix}$$

- If we put the two column vectors x and y in an $n \times 2$ matrix $Z = [x, y]$, then $M^0 Z$ is the $n \times 2$ matrix in which the two columns of data are in mean deviation form.

- A **vector space** is any set of vectors that is closed under scalar multiplication and addition. A 2(or K -)-dimensional Euclidean space is called R^2 or (R^K) .
- A set of vectors in a vector space is a **basis** for that vector space if any vector in the space can be written as a linear combination of them.
- A set of vectors is **linearly dependent** if any one of the vectors in the set can be written as a linear combination of the others.
- A set of vectors, $a_1 \dots, a_K$, is **linearly independent** *iff* the only solution to

$$\lambda_1 a_1 + \dots + \lambda_K a_K = 0$$

is

$$\lambda_1 = \dots = \lambda_K = 0$$

- A basis for a vector space of K dimensions is any set of K linearly independent vectors in that space.
- The set of all linear combinations of a set of vectors is the vector space that is **spanned** by those vectors.
- The **column space** of a matrix is the vector space that is spanned by its column vectors.
- The **column rank** of a matrix is the dimension of the vector space that is spanned by its column. Similarly one can define the row rank.

- The column rank is always equal to the row rank, which is called the **rank** of the matrix..
- A matrix is of **full column rank** if the column rank equals the number of columns. Similarly one can define the full row rank.
- In this course, full rank means full column rank.
- $\text{rank}(A)=\text{rank}(A')\leq\min(\text{number of rows, number of columns})$
- $\text{rank}(AB)\leq\min(\text{rank}(A), \text{rank}(B))$
- $\text{rank}(A)=\text{rank}(A')=\text{rank}(AA')$
- Determinant of a 2×2 square matrix. $\begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad - bc$
- $|A| = \sum_{j=1}^K a_{ij}(-1)^{i+j}|A_{ij}|$, where A_{ij} is the matrix obtained from A by deleting row i and column j and its determinant is called a **minor** of A . $(-1)^{i+j}|A_{ij}|$ is called a **cofactor**.
- $|A| = |A'|$.

- Suppose e is a column vector. The **length** or **norm** of it is $\|e\| = \sqrt{e'e}$.
- Suppose a and b are two column vectors. They are **orthogonal** ($a \perp b$) iff $a'b = b'a = 0$.
- Suppose a and b are two column vectors. The angle θ between these two vectors satisfies $\cos \theta = \frac{a'b}{\|a\| \cdot \|b\|}$.
- Suppose y is an n -dimensional column vector, X is an $n \times K$ (assuming $n \geq K$) matrix (hence can be partitioned into K column vectors, each with n dimensions). If y is in the column space of X , then one can find an K -dimensional column vector, b , so that $y = Xb$.
- Suppose y is NOT in the column space of X , then no column vector, b , satisfies $y = Xb$. Indeed for any b , $y - Xb = e$ will be different from 0. Of course, for any b , Xb must be in the column space of X .
- Least squares problem: find the b so that $\|e\| = \|y - Xb\|$ is the smallest. The solution is to make e orthogonal to Xb , ie, $(Xb)'e = 0$. That implies $X'y = X'Xb$.
- Suppose A is a square matrix. If exists a square matrix B such that $BA = I$. We call B the **inverse** of A and denote it by A^{-1} .
- A matrix whose inverse exists is **nonsingular**. In this case $|A| \neq 0$ and its rank equals the number of columns (also rows).

- Suppose a^{ij} is the ij th element of A^{-1} . The general formula for computing an inverse matrix is $a^{ji} = \frac{|C_{ij}|}{|A|}$, where $|C_{ij}|$ is the ij th **co-factor** of A .
- Properties about the inverse:
 1. $|A^{-1}| = |A|^{-1}$
 2. $(A^{-1})^{-1} = A$
 3. $(A^{-1})' = (A')^{-1}$
 4. $(AB)^{-1} = B^{-1}A^{-1}$ where A, B are both square matrices.

- $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ is a **partitioned** matrix, where normally A_{11} and A_{22} are square matrices. If A_{21} and A_{12} are 0, A is a **block diagonal** matrix.

- $\begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix} = |A_{22}| \cdot |A_{11} - A_{12}A_{22}^{-1}A_{21}| = |A_{11}| \cdot |A_{22} - A_{21}A_{11}^{-1}A_{12}|$

- $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}^{-1} = \begin{bmatrix} A_{11}^{-1}(I + A_{12}F_2A_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}F_2 \\ -F_2A_{21}A_{11}^{-1} & F_2 \end{bmatrix},$

where $F_2 = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}$.

- An example: suppose x is a n -dimensional column vector.

$$A = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} i' & i'x \\ x'i & x'x \end{bmatrix}.$$

Then $F_2 = (x'x - x'i(i'i)^{-1}i'x)^{-1} = (x'M^0x)^{-1}$

- **Kronecker product:** $A \otimes B = [a_{ij}B]$.
- Properties about the Kronecker product: $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$,
 $(A \otimes B)' = A' \otimes B'$, $(A \otimes B)(C \otimes D) = AC \otimes BD$
- Suppose A is a square matrix, $c \neq 0$ is a column vector and λ is a scalar. If $Ac = \lambda c$ is satisfied, then we call c a **characteristic vector (eigenvector)** of A and λ a **characteristic root (eigenvalue)** of A . We often normalize c so that the norm is 1.
- The characteristic roots of a symmetric matrix are real. To find the characteristic roots of A , one needs to solve $|A - \lambda I| = 0$.
- Suppose the characteristic roots of symmetric matrix A are $\lambda_1 \dots, \lambda_K$. The corresponding characteristic vectors are c_1, \dots, c_K . Then $Ac_k = \lambda_k c_k$
- If $\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_K \end{bmatrix} \equiv \text{diag}(\lambda_1 \dots, \lambda_K)$, then $AC = C\Lambda$, where $C = [c_1 \ c_2 \ \dots \ c_K]$ that satisfies $C' = C^{-1}$ and $CC' = CC^{-1} = I$.
- **Diagonalization** of a matrix: $C'AC = \Lambda$. Spectral decomposition of A : $A = C\Lambda C'$.
- The rank of a matrix is the number of nonzero characteristic roots it contains.
- Condition number $\gamma = \left[\frac{\text{maximum root}}{\text{minimum root}} \right]^{1/2}$. Matrices with large condition numbers are difficult to invert accurately.

- **Trace** of a matrix: $tr(A) = \sum a_{kk}$ which equals the sum of its characteristic roots.
 1. $tr(cA) = ctr(A)$
 2. $tr(A') = tr(A)$
 3. $tr(A + B) = tr(A) + tr(B)$
 4. $tr(AB) = tr(BA)$
 5. $tr(ABC) = tr(BCA) = tr(CAB)$

- The determinant of a matrix equals the product of its characteristic roots.

- The characteristic roots of A^2 are the squares of those of A .

- Define $A^{1/2}$ to be a matrix such that $A^{1/2}A^{1/2} = A$. The characteristic roots of $A^{1/2}$ are the square roots of those of A .

- Factorization of a matrix (ie find a P such that $P'P = A^{-1}$):
 1. One choice is $P = \Lambda^{-1/2}C'$.
 2. **Cholesky factorization:** $A = LU$, where L is a lower triangular matrix and U is an upper triangular matrix.

- Quadratic form: $q = \sum_{i=1}^n \sum_{j=1}^n x_i x_j a_{ij} = x'Ax$.

- If $x'Ax > 0$ for all nonzero x , A is **positive definite**. We sometimes write $A > 0$.

- If $x'Ax < 0$ for all nonzero x , A is **negative definite**.
- If $x'Ax \geq 0$ for all nonzero x , A is **positive semidefinite**. We sometimes write $A \geq 0$.
- Let A be a symmetric matrix. If all the characteristic roots of A are positive (negative), then A is positive (negative) definite.
- The only full rank, symmetric idempotent matrix is the identity matrix. All the other symmetric idempotent matrices are singular.
- Every symmetric idempotent matrix is positive semidefinite with roots being either one or zeros.

2 Calculus

- A function $f(x)$ is continuous at x_0 if $\lim_{x \rightarrow x_0} f(x) = f(x_0)$. It is continuous at an interval I if it is continuous at every point in I .
- A sequence of functions, $\{f_n(x)\}$ is said to be uniformly convergent to $f(x)$ on I if, every $\varepsilon > 0$, there exists N , such that $|f_n(x) - f(x)| < \varepsilon$ for all $n > N$ and for all $x \in I$.
- Let $\{f_n(x)\}$ be a sequence of functions on I . The sequence $\{f_n(x)\}$ is equicontinuous if for every $\varepsilon > 0$ and every $x \in I$, there exists a $\delta > 0$, such that for all n and all $x' \in X$ with $|x' - x| < \delta$ we have $|f_n(x) - f_n(x')| < \varepsilon$.
- The sequence $\{f_n(x)\}$ is uniformly equicontinuous if for every $\varepsilon > 0$, there exists a $\delta > 0$, such that for all n and all $x, x' \in I$ with $|x' - x| < \delta$ we have $|f_n(x) - f_n(x')| < \varepsilon$.
- For comparison, the statement all functions $f_n(x)$ are continuous means that for every $\varepsilon > 0$, every n , and every $x \in I$, there exists a $\delta > 0$, such that for all $x' \in I$ with $|x' - x| < \delta$ we have $|f_n(x) - f_n(x')| < \varepsilon$. So, for continuity, δ may depend on ε , x and n ; for equicontinuity, δ must be independent of n ; and for uniform equicontinuity, δ must be independent of both n and x .
- Let $\{f_n(x)\}$ be an equicontinuous sequence of functions. If $f_n(x) \rightarrow f(x)$ for every $x \in I$, then the function f is continuous.

- Let $\{f_n(x)\}$ uniformly converges to $f(x)$ and each $f_n(x)$ is continuous on I , then the function f is continuous.
- Every equicontinuous sequence of functions from $[0, 1]$ to R is uniformly equicontinuous.
- Let $\{f_n(x)\}$ be an equicontinuous sequence of functions from $[0, 1]$ to R . If $f_n(x) \rightarrow f(x)$ for every $x \in [0, 1]$, then $f_n(x) \rightarrow f(x)$ uniformly in x .
- Let $\{f_n(x)\}$ be an equicontinuous sequence of uniformly bounded functions from $[0, 1]$ to R . Then there is a subsequence which converges uniformly.

- A univariate Taylor series expansion: $f(x) \approx f(x_0) + \sum_{i=1}^p \frac{1}{i!} \frac{d^i f(x_0)}{d(x_0)^i} (x - x_0)^i$
- Now suppose $f : R^n \rightarrow R^1$. The gradient (Jacobian) is $\frac{\partial f(x)}{\partial x} = g = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$. The Hessian is $\frac{\partial^2 f(x)}{\partial x \partial x'} = H = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$
- The second order Taylor series expansion: $f(x) \approx f(x_0) + (g^0)'(x - x_0) + \frac{1}{2}(x - x_0)'H^0(x - x_0)$
- Now suppose $f : R^n \rightarrow R^K$. The Jacobian is ∇f and the Hessian is $\nabla^2 f$. Then second order Taylor series expansion is $f(x) \approx f(x_0) + \nabla f(x_0)(x - x_0) + \frac{1}{2}\nabla^2 f(x_0)[(x - x_0) \otimes (x - x_0)]$.
- A linear function can be written as $y = a'x = \sum_{i=1}^n a_i x_i$. A quadratic function can be written as $y = x'Ax = \sum_{i=1}^n \sum_{j=1}^n x_i x_j a_{ij}$. Then
 - $\frac{\partial(a'x)}{\partial x} = \frac{\partial(x'a)}{\partial x} = a$
 - $\frac{\partial(x'Ax)}{\partial x} = 2Ax$ when A is symmetric.
- Optimization: $\max f(x)$ or $\min f(x)$. Suppose $f : R^1 \rightarrow R^1$ is differentiable and the optimum takes place at an interior point.
- First order (necessary condition) condition: $\frac{df(x)}{dx} = 0$
- The sufficient condition is: $\frac{d^2 f(x)}{dx^2} < 0$ for a maximum and $\frac{d^2 f(x)}{dx^2} > 0$ for a minimum.

- Now suppose $f : R^n \rightarrow R^1$. First order condition is: $\frac{\partial f(x)}{\partial x} = 0$. The sufficient condition is: H is negative definite for a maximum and positive definite for a minimum.

3 Review of Statistics

- Classical definition of probability:

Experiment

↓

Outcomes

↓

Sample Space

↓

Event

- Probability of an event:

$$P(A) = \frac{\text{Total number of outcomes favorable to A}}{\text{Total number of outcomes}}$$

- Properties:

- $P(\Phi) = 0$, where Φ is an empty space.

- $P(\Omega) = 1$

- For any A , $0 \leq P(A) \leq 1$

- $P(A \cup B) = P(A) + P(B)$ if A and B disjoint

- $P(\bar{A}) = 1 - P(A)$

- $A \subset B \Rightarrow P(A) \leq P(B)$

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- More definitions:

- Conditional Probability

A, B are two events and $P(B) \neq 0$, then the conditional probability of A conditional on B is defined as

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

- A and B are independent if

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

- Random variable:

A random variable is a rule which assigns a numerical value to any outcome of the experiment. X, Y, Z (capital letters) are often used to represent a random variable.

- Distribution function(**cumulative distribution function: cdf**)

The distribution function of a r.v. is the function $F : \mathcal{R} \rightarrow [0, 1]$ and is given by $F(x) = P(X \leq x)$

- A random variable is discrete if it can take only a finite or countable number of values in \mathcal{R} . When a r.v. is discrete, $f(x) = P(X = x)$ is called the probability mass function.

- A random variable is continuous if $\exists f(x)$ such that $F(x) = \int_{-\infty}^x f(u)du$, where $f(x)$ is called the **probability density function** (pdf).

- $S(x) = 1 - F(x)$ is called the **survival function**. $h(x) = \frac{f(x)}{S(x)}$ is called the **harzard function**.

- For a random variable X , if the function $M(t) = E(e^{tX})$ exists, then it is the **moment-generating function**.

- For a random variable X , $c(t) = E(e^{itX})$ is called the **characteristic function**, where $i = \sqrt{-1}$. $\psi(t) = \ln c(t)$ is called the **cumulant generating function**.

• Properties:

$$- \lim_{x \rightarrow \pm\infty} F(x) = \begin{cases} 1 \\ 0 \end{cases}$$

$$- x < y \Rightarrow F(x) \leq F(y)$$

$$- P(X > x) = 1 - F(x)$$

$$- P(x < X \leq y) = F(y) - F(x)$$

$$- P(X = x) = \lim_{h \searrow 0} P(x - h < X \leq x) = F(x) - \lim_{h \searrow 0} F(x - h)$$

For the discrete random variable

$$- F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

$$- \sum_{x_i} f(x_i) = 1$$

For the continuous random variable

$$- F'(x) = f(x)$$

$$- P(X = x) = F(x) - \lim_{h \searrow 0} F(x - h) = \lim_{h \searrow 0} \int_{x-h}^x f(u) du = 0$$

$$- \int_{-\infty}^{\infty} f(u) du = 1$$

$$- P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = F(b) - F(a) = \int_a^b f(u) du$$

- Sometimes it is hard to know the pdf or cdf of a random variable. In other cases, it is not necessary to know the whole pdf or cdf. We may only need to know some important characteristics of it.

- Definitions:

- Expectation or expected value or mean:

$$E(X) = \begin{cases} \sum xf(x) & \text{if } X \text{ is a discrete random variable.} \\ \int_{-\infty}^{\infty} xf(x)dx & \text{if } X \text{ is a continuous random variable.} \end{cases}$$

More general,

$$E(u(X)) = \begin{cases} \sum u(x)f(x) & \text{if } X \text{ is a discrete random variable.} \\ \int_{-\infty}^{\infty} u(x)f(x)dx & \text{if } X \text{ is a continuous random variable.} \end{cases}$$

- Variance:

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

- $\sigma = \sqrt{Var(X)}$ is called the standard deviation of X

- A number m is the τ th quantile of a random variable X if $Prob(X \geq m) = 1 - \tau$ and $P(X < m) = \tau$. When $\tau = 0.5$, m is the median.

- skewness= $E[(X - \mu)^3]$

- skewness coefficient= $\frac{E[(X-\mu)^3]}{\sigma^3}$

- kurtosis= $E[(X - \mu)^4]$

- kurtosis for Normal distribution is 3.

- degree of excess kurtosis= $\frac{E[(X-\mu)^4]}{\sigma^4} - 3$.

- Properties: (a, b, c are all constants)

- $E(c) = c$

- $E(aX + bY) = aE(X) + bE(Y)$

- $Var(X) = EX^2 - (E(X))^2$

- $Var(c) = 0$

- $Var(aX + b) = a^2Var(X)$

- It can be shown that $\frac{d^k M(t)}{dt^k} \Big|_{t=0} = E(X^k)$ (**moments**)

- Definitions:

- Joint distribution:

The joint cdf and joint pdf for random variables X and Y are defined as

$$F(x, y) = P(X \leq x, Y \leq y) = \begin{cases} \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j) & \text{if } X, Y \text{ discrete} \\ \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv & \text{if } X, Y \text{ continuous} \end{cases}$$

where $F(x, y)$ is the joint cdf and $f(x, y)$ is the joint pdf

- Marginal distribution:

$f_1(x)$ and $f_2(y)$ are marginal pdfs and defined by

$$f_1(x) = \sum_y f(x, y), \quad f_2(y) = \sum_x f(x, y)$$

- Conditional distribution:

$f(x|y)$ and $f(y|x)$ are marginal pdfs and defined by

$$f(x|y) = \frac{f(x, y)}{f_2(y)}, \quad f(y|x) = \frac{f(x, y)}{f_1(x)}$$

- $Y|X = x$ is a random variable. $E(Y|X = x) = \int_{-\infty}^{+\infty} y f(y|x) dy$ is called the conditional expectation. It is a real function of x .

– $E(Y|X)$ is a random variable. $E_X(E_Y(Y|X)) = E(Y)$. This is known as the **Law of Iterative Expectations**.

– $Var(Y) = Var_X(E_Y(Y|X)) + E_X(Var_Y(Y|X))$. This is known as the **Decomposition of Variance**.

– Independence of two random variables:

The random variables X and Y are independent iff $f(x, y) = f_1(x)f_2(y)$

– Covariance between two random variables:

$$\sigma_{X,Y}^2 = Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

$$\text{where } E(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x, y)dx dy$$

– Covariance matrix (or variance-covariance matrix):

$$\Sigma_{X,Y} = \begin{pmatrix} \sigma_X^2 & Cov(X, Y) \\ Cov(X, Y) & \sigma_Y^2 \end{pmatrix}$$

– Correlation coefficient:

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \in [-1, 1]$$

When $\rho_{X,Y} = 0$, we say X and Y are uncorrelated.

- Properties:
 - If X and Y are independent, then $f(X)$ and $g(Y)$ are independent, where f and g are any measurable functions.
 - If X and Y are independent, then $E(XY) = E(X)E(Y)$
 - If X and Y are independent, then $Cov(X, Y) = 0$. **But not vice versa!**
 - $Cov(X, X) = Var(X)$, $\rho_{X,X} = 1$
 - $Var(aX + bY) = a^2Var(X) + b^2Var(Y) + 2abCov(X, Y)$
 - if X and Y are uncorrelated, then $Var(aX + bY) = a^2Var(X) + b^2Var(Y)$
- Extension of 2×2 covariance matrix to high dimensional matrix:

We have T different random variables, X_1, \dots, X_T

$$\Sigma = \begin{pmatrix} \sigma_{X_1}^2 & Cov(X_1, X_2) & \cdots & Cov(X_1, X_T) \\ Cov(X_2, X_1) & \sigma_{X_2}^2 & \cdots & Cov(X_2, X_T) \\ \vdots & \vdots & \cdots & \vdots \\ Cov(X_T, X_1) & Cov(X_T, X_2) & \cdots & \sigma_{X_T}^2 \end{pmatrix}$$

- Some important distribution functions:

1. Normal distribution

- Normal distribution: $N(\mu, \sigma^2)$

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

A normal distribution is uniquely determined by μ, σ^2

$N(0,1)$ is called the standard normal distribution

- Multivariate normal distribution: $N(\boldsymbol{\mu}, \Sigma)$

$$f(\mathbf{x}) = (2\pi)^{-\frac{T}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

- Properties

(a) $X \sim N(\mu, \sigma^2) \Rightarrow E(X) = \mu, Var(X) = \sigma^2$

(b) $X \sim N(\mu, \sigma^2) \Rightarrow Z = (X - \mu)/\sigma \sim N(0, 1)$

(c) $X \sim N(\mu, \sigma^2) \Rightarrow aX + b \sim N(a\mu + b, a^2\sigma^2)$

(d) $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2) \Rightarrow X + Y \sim N(\mu_1 + \mu_2, Var(X + Y))$

(e) Moment generating function for $N(\mu, \sigma^2)$: $\exp(\mu t + \frac{1}{2}\sigma^2 t^2)$

(f) Moment generating function for $N(0, 1)$: $\exp(\frac{1}{2}t^2)$

(g) Charateristic function for $N(\mu, \sigma^2)$: $\exp(i\mu t - \frac{1}{2}\sigma^2 t^2)$

- (h) $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{X,Y}^2 \\ \sigma_{X,Y}^2 & \sigma_Y^2 \end{pmatrix} \right]$
 $\Rightarrow X|Y = y$ and $Y|X = x$ are both normal and
 $X|Y = y \sim N \left(\mu_1 + \frac{\sigma_{X,Y}^2(y - \mu_2)}{\sigma_Y^2}, \sigma_X^2 - \frac{\sigma_{X,Y}^4}{\sigma_Y^2} \right)$
 $Y|X = x \sim N \left(\mu_2 + \frac{\sigma_{X,Y}^2(x - \mu_1)}{\sigma_X^2}, \sigma_Y^2 - \frac{\sigma_{X,Y}^4}{\sigma_X^2} \right)$
- (i) If $\sigma_{X,Y} = 0$, then X and Y are independent

2. χ^2 distribution:

- $\chi_{(1)}^2 \equiv Z^2$ where $Z \sim N(0, 1)$
- If X_1, \dots, X_r are independently and identically distributed (IID) as $\chi_{(1)}^2$ distribution, then
 $Y = X_1 + \dots + X_r \sim \chi_{(r)}^2$
- If $Y \sim \chi_{(r)}^2$, $E(Y) = r$, $Var(Y) = 2r$
- Suppose $X \sim N(0, I)$ and A is idempotent, then $X'AX$ has a χ^2 distribution with degrees of freedom being equal to the ranks of A .
- Suppose $X \sim N(0, I)$ and A and B are idempotent with $AB = 0$. Then $X'AX$ and $X'BX$ are independent.

3. t distribution:

– $t_{(T)} = \frac{Z}{\sqrt{\frac{Y}{T}}}$, where $Y \sim \chi_{(T)}^2$ and independent of Z which is $N(0, 1)$

– $E(t_{(T)}) = 0$, $Var(t_{(T)}) = \frac{T}{T-2}$

4. Cauchy distribution

– Cauchy distribution is obtained from $\frac{X}{Y}$, where $X \sim N(0, 1)$, $Y \sim N(0, 1)$ and they are independent

– Mean and variance do not exist. Median is 0

5. F distribution:

– $F_{(r_1, r_2)} = \frac{X/r_1}{Y/r_2}$, where $X \sim \chi_{(r_1)}^2$, $Y \sim \chi_{(r_2)}^2$ and they are independent

6. Lognormal, gamma, beta, logistic

7. Poisson, Binomial

• Properties:

- If $X \sim N(0, I)$ then $X'X \sim \chi_{(T)}^2$
- If $X \sim N(\mu, \Sigma)$ then $L'X \sim N(L'\mu, L'\Sigma L)$
- If $X \sim N(\mu, \Sigma)$, then $(X - \mu)'\Sigma^{-1}(X - \mu) \sim \chi_{(T)}^2$
- If $X \sim N(\mu, \Sigma)$, then $\Sigma^{-1/2}(X - \mu) \sim N(0, I)$
- Suppose $X \sim N(0, I)$, A is idempotent and L is a matrix with $LA = 0$. Then LX and $X'AX$ are independent.
- Let $d = \mu + \sigma^2 \frac{\partial}{\partial \mu}$. If $f(y)$ is the pdf of $N(\mu, \sigma^2)$, then $(y - d)f(y) = 0$ or $yf(y) = df(y)$.
- More generally, if h a real-valued function of y , then $(h(y) - h(d))f(y) = 0$ or $h(y)f(y) = h(d)f(y)$.
- If $y \sim N(\mu, \sigma^2)$, $E(h(y))$ exists, then $E(h(y)g(y)) = h(d)E(g(y))$

- **Change of variable technique.** Suppose X is a random variable whose probability density function (pdf) is $f_X(x)$. Let g be a differentiable and monotonic function. What is the pdf of $Y = g(X)$ (call it $f_Y(y)$)?
- Let $h(x) = g^{-1}(x)$ be the inverse of $g(x)$, ie, $h(g(x)) = x$.
- Then the change of variable technique says that

$$f_Y(y) = f_X(h(y))|h'(y)|.$$

- Definitions:
 1. A parameter is a number that describes the population characteristics
 2. An estimate is a value calculated from a sample of data that is used to estimate a population parameter. It is a number, not a random variable
 3. The formula to calculate an estimate is called an estimator. It is a random variable and varies with samples.
 4. A sample statistic is a rule to tell us how to describe some characteristics of a sample
 5. The sampling distribution is the probability distribution of any sample statistic

- Example:

- Population: A random variable Y represents the household income in New Zealand.

- Parameters: Usually they have important economic meaning.

$\mu = E(Y)$ – average household income

$\sigma^2 = Var(Y)$ – household income dispersion (inequality)

- In general the parameters are unknown to econometricians. We have to estimate them from a dataset. Before we collect a dataset (sample), we need to decide a rule to estimate the parameters. Suppose $\{Y_1, \dots, Y_T\}$ is a sample.

- Estimator:

For μ , $\bar{Y} = \frac{Y_1 + \dots + Y_T}{T}$ can be an estimator. It is called the sample mean.

For σ^2 , $S_Y^2 = \frac{(Y_1 - \bar{Y})^2 + \dots + (Y_T - \bar{Y})^2}{T-1}$ can be an estimator. It is called the sample variance.

Both \bar{Y} and S_Y^2 are random variables.

– Estimate:

Suppose we actually collect a sample $\{y_1, \dots, y_T\}$. Since y_1, \dots, y_T are real numbers, they are not random variables. $\frac{y_1 + \dots + y_T}{T}$ is an estimate of μ . It is called the point estimate. The quality of the estimation depends on both the sample and the sample size.

– Random sample:

$\{Y_1, \dots, Y_T\}$ is a random sample from a population Y if

1. Y_1, \dots, Y_T have the same distribution function as Y
2. Y_1, \dots, Y_T are independent

– Sampling distribution:

If $\{Y_1, \dots, Y_T\}$ is a random sample from a population $Y \sim N(\mu, \sigma^2)$.

Then $\bar{Y} \sim N(\mu, \sigma^2/T)$ and $\frac{(T-1)S_Y^2}{\sigma^2} \sim \chi_{(T-1)}^2$

- Criteria to evaluate an estimator
 - Unbiasedness: θ is an unbiased estimator of β if $E(\theta) = \beta$
 - Efficiency: If both θ_1 and θ_2 are both unbiased estimators of β , and $Var(\theta_1) < Var(\theta_2)$, then θ_1 is more efficient than θ_2 .
 - Weak consistency: Suppose θ_T is an estimator of β . If $\forall \varepsilon > 0, \lim_{T \rightarrow \infty} P(|\theta_T - \beta| < \varepsilon) = 1$, then θ_T is a consistent estimator of β in the weak sense. Usually we denote it by $\text{plim} \theta_T = \beta$ or $\theta_T \xrightarrow{p} \beta$
 - Strong consistency: Suppose θ_T is an estimator of β . If $P(\lim_{T \rightarrow \infty} \theta_T = \beta) = 1$, then θ_T is a consistent estimator of β in the strong sense. Usually we denote it by $\theta_T \xrightarrow{as} \beta$

- Form of an estimator

Linear estimator is a linear function of the sample observations. An estimator which is linear and also has the smallest variance among all linear unbiased estimators is called the best linear unbiased estimator (BLUE).

- Properties of sample mean:
 1. Unbiasedness
 2. Linearity
 3. BLUE
 4. Consistency: Weak Law of Large Number (LLN), ie, $\text{plim}\bar{Y} = \mu$.

- From the inference point of view, point estimation is not enough. We have to know something about the distribution of the estimator(sampling distribution). For example, if σ^2 is known, the sampling distribution of \bar{Y} is

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{T}\right)$$

The sampling distribution is very important for inferences, such as constructing confidence intervals, testing hypothesis and making prediction

- Confidence interval of μ :

$$(\bar{Y} - Z_{\alpha/2} \frac{\sigma}{\sqrt{T}}, \bar{Y} + Z_{\alpha/2} \frac{\sigma}{\sqrt{T}})$$

where $Z_{\alpha/2}$ is the critical value and $1 - \alpha$ is the level of confidence

- Hypothesis testing:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

- If σ^2 is known, we can use $\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{T}}$ to test the hypothesis. The sampling distribution of $\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{T}}$ is $N(0, 1)$. If $|\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{T}}| > Z_{\alpha/2}$ we reject H_0 at α significant level, where $Z_{\alpha/2}$ is the critical value.

Otherwise, we will not reject H_0 .

- If σ^2 is unknown, we can not use $\frac{\bar{Y} - \mu_0}{\sigma/\sqrt{T}}$ to test the hypothesis. Instead we define $\hat{\sigma} = \sqrt{S_Y^2}$ and a new test statistic $\frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{T}}$. The sampling distribution of $\frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{T}}$ is not $N(0, 1)$. In fact $\frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{T}} \sim t_{(T-1)}$. If $|\frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{T}}| > t_{T-1, \alpha/2}$ we reject H_0 at α significant level, where $t_{T-1, \alpha/2}$ is the critical value. Otherwise, we will not reject H_0 . The test statistic $\frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{T}}$ is called the t-statistic.

- Test for normality.

Single Tests

A simple way to test for normality is to compare the computed skewness and kurtosis coefficients with the theoretical values under the assumption of normality; namely 0 and 3 respectively. Thus, the tests are

$$\text{Skewness Test : } Z_{Sk} = \frac{S}{\sqrt{6/T}},$$

$$\text{Kurtosis Test : } Z_{Kt} = \frac{K - 3}{\sqrt{24/T}}.$$

Both test statistics are distributed under the null hypothesis of normality as $N(0, 1)$. Thus “large” values of the test statistic, say greater than ± 2 , constitute rejection of the null hypothesis of normality.

Joint Test (JB test)

A joint test can also be constructed as

$$JB = Z_{Sk}^2 + Z_{Kt}^2.$$

which is distributed as χ_2^2 . The null hypothesis of normality is rejected at the 5% level when the p-value is less than $\alpha = 0.05$. This test is commonly referred to as the Jarque-Bera test of normality.

- What happens if $Y \approx \text{Normal}$

In this case \bar{Y} is not exactly normally distributed. Fortunately, \bar{Y} is approximately normally distributed when the sample size T is reasonably large.

- Central Limit Theorem (CLT)

If $Y \sim (\mu, \sigma^2)$ (but not necessarily $N(\mu, \sigma^2)$) and $\{Y_1, \dots, Y_T\}$ is a random sample. Then \bar{Y} is approximately normally distributed when the sample size T is large enough. In fact,

$$\sqrt{T}(\bar{Y} - \mu) \xrightarrow{d} N(0, \sigma^2) \quad (\text{or briefly, } \bar{Y} \overset{a}{\sim} N(\mu, \sigma^2/T))$$

- Most econometric inferences rely on either the finite sample distribution (when it is available) or the asymptotic distribution.

- Definitions:

1. Type I error = $\Pr(\text{reject } H_0 | H_0 \text{ is true}) = \alpha$. This is the size of the test.
2. Type II error = $\Pr(\text{accept } H_0 | H_0 \text{ is false}) = \beta$
3. power of the test: $1 - \beta$
4. p-value: the lowest significant level at which H_0 can be rejected.
5. A test is most powerful if it has greater power than any test of the same size.
6. A test is consistent if the power goes to one as the sample size grows to infinity.

4 Asymptotic Theory

- **Converge in probability:** $\forall \varepsilon > 0, \lim_{n \rightarrow \infty} P(|x_n - c| < \varepsilon) = 1$.
Usually we denote it by $\text{plim} x_n = c$ or $x_n \xrightarrow{p} c$.
- **Markov inequality:** suppose y_n is a nonnegative random variable,
 $\forall \delta > 0, P(y_n \geq \delta) \leq E(y_n)/\delta$.
- **Chebychev's inequality:** suppose y_n is a random variable, $\forall c, \varepsilon$,
 $P(|x_n - c| > \varepsilon) \leq E((x_n - c)^2)/\varepsilon^2$.
- Suppose x_n has mean μ_n and σ_n^2 such that the ordinary limit of them are c and 0 , respectively. Then we say x_n **converges in mean square** to c or $x_n \xrightarrow{r^2} c$.
- If $x_n \xrightarrow{r^2} c$ then $x_n \xrightarrow{p} c$.
- **Weak Law of Large Numbers:** Suppose x_1, \dots, x_n are a sequence of independent and identically distributed (iid) random variables with mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Then $\bar{x} \xrightarrow{p} \mu$.
- **Khinchine's Weak Law of Large Numbers:** Suppose x_1, \dots, x_n are a sequence of iid random variables with mean $\mu < \infty$. Then $\bar{x} \xrightarrow{p} \mu$.
- **Chebychev's Weak Law of Large Numbers:** Suppose x_1, \dots, x_n are a sequence of independent random variables with $E(x_i) = \mu_i < \infty$ and $\text{Var}(x_i) = \sigma_i^2 < \infty$ such that $\frac{1}{n} \bar{\sigma}_n^2 = \frac{1}{n} \left(\frac{1}{n} \sum \sigma_i^2 \right) \rightarrow 0$ as $n \rightarrow \infty$. Then $\bar{x} - \frac{1}{n} \sum \mu_i \xrightarrow{p} 0$.

- **Almost Sure Convergence:** $P(\lim_{n \rightarrow \infty} x_n = c) = 1$. Usually we denote it by $x_n \xrightarrow{a.s.} c$.
- If $x_n \xrightarrow{a.s.} c$ then $x_n \xrightarrow{p} c$.
- **Kolmogorov's Strong Law of Large Numbers:** Suppose x_1, \dots, x_n are a sequence of iid random variables with $E(x_i) = \mu < \infty$ and $E(|x_i|) < \infty$. Then $\bar{x} - \mu \xrightarrow{a.s.} 0$.
- **Kolmogorov's Strong Law of Large Numbers:** Suppose x_1, \dots, x_n are a sequence of independent random variables with $E(x_i) = \mu_i < \infty$ and $Var(x_i) = \sigma_i^2 < \infty$ such that $\sum \sigma_i^2 / i^2 < \infty$ as $n \rightarrow \infty$. Then $\bar{x} - \frac{1}{n} \sum \mu_i \xrightarrow{a.s.} 0$.
- **Markov's Strong Law of Large Numbers:** Suppose x_1, \dots, x_n are a sequence of independent random variables with $E(x_i) = \mu_i < \infty$. If for some $\delta > 0$, $\sum E(|x_i - \mu_i|^{1+\delta}) / i^{1+\delta} < \infty$ as $n \rightarrow \infty$. Then $\bar{x} - \frac{1}{n} \sum \mu_i \xrightarrow{a.s.} 0$.
- **Slutsky Theorem:** $\text{plim}g(x_n) = g(\text{plim}x_n)$ where g is a continuous function.
- **Convergence in Distribution:** Suppose x_1, \dots, x_n are a sequence of random variables with cdf $F_n(x_n)$. Suppose x is a random variable with cdf $F(x)$. If $F_n(x_n) \rightarrow F(x)$ for all continuity point of F , we say $x_n \xrightarrow{d} x$. The mean (variance) of x is called the limiting mean (variance).

- If $x_n \xrightarrow{d} x$ then $g(x_n) \xrightarrow{d} g(x)$ where g is a continuous function.
- If $x_n \xrightarrow{d} x$, $y_n \xrightarrow{p} a$, $z_n \xrightarrow{p} b$, then $x_n y_n + z_n \xrightarrow{d} ax + b$
- **Cramer-Wold Device:** If $x_n \xrightarrow{d} x$, then $c'x_n \xrightarrow{d} c'x$

- **Lindberg-Levy Central Limit Theorem:** Suppose x_1, \dots, x_n are a sequence of iid random variables with mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Then $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2)$.
- **Lindberg-Feller Central Limit Theorem:** Suppose x_1, \dots, x_n are a sequence of independent random variables with $E(x_i) = \mu_i < \infty$ and $Var(x_i) = \sigma_i^2 < \infty$ such that $\lim \max_{i \in [1, n]} (\sigma_i) / (n\bar{\sigma}_n) = 0$ as $n \rightarrow \infty$. Then $\sqrt{n}(\bar{x} - \bar{\mu}_n) \xrightarrow{d} N(0, \lim \bar{\sigma}_n^2)$, where $\bar{\sigma}_n^2 = \frac{1}{n} \sum \sigma_i^2$.
- **Liapounov Central Limit Theorem:** Suppose x_1, \dots, x_n are a sequence of independent random variables with $E(x_i) = \mu_i < \infty$ and $Var(x_i) = \sigma_i^2 < \infty$ such that $E[|x_i - \mu_i|^{2+\delta}]$ is finite for some $\delta > 0$. Then $\sqrt{n}(\bar{x} - \bar{\mu}_n) / \bar{\sigma}_n \xrightarrow{d} N(0, 1)$ if $\bar{\sigma}_n > 0$.
- **Delta method:** if $\sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N(0, \sigma^2)$ and if g is a differentiable function, then $\sqrt{n}(g(\bar{x}) - g(\mu)) \xrightarrow{d} N\left(0, \frac{dg(\mu)}{d\mu} \sigma^2 \frac{dg(\mu)}{d\mu}\right)$
- The random sequence $\{b_n\}$ is at most of order n^λ in probability, denoted $b_n = O_p(n^\lambda)$, if for every $\varepsilon > 0$ there exists a finite $\Delta(\varepsilon) > 0$ and $N(\varepsilon) \in \mathbb{N}$, such that $P(|n^{-\lambda}b_n| > \Delta(\varepsilon)) < \varepsilon$ for all $n \geq N(\varepsilon)$.
- The random sequence $\{b_n\}$ is of smaller order n^λ in probability, denoted $b_n = o_p(n^\lambda)$, if $n^{-\lambda}b_n \xrightarrow{p} 0$.

5 Computation

- Random draw from univariate and multivariate distributions. How to generate a random draw from a distribution F ? Define F^{-1} be the inverse of F . Let $X \sim U[0, 1]$. Then $Y = F^{-1}(X) \sim F$.
- Gibbs sampler generates a correlated draw. Example: How to draw from $f(x_1, x_2)$? Repeated draws from $f(x_2|x_1)$ and then from $f(x_1|x_2)$.
- Numerical integration: Simpson's rule, Gaussian quadrature, Monte Carlo methods. Example: Suppose $f : [0, 1] \rightarrow [0, 1]$. We need to evaluate $\int_0^1 f(x)dx$. One form of Monte Carlo method is to approximate $\int_0^1 f(x)dx$ by $\frac{1}{n} \sum_{i=1}^n f(x_i)$ where $x_i \sim U[0, 1]$.
- Variance reduction methods: antithetic variable, importance sampling
- Numerical optimization: grid search, Newton's method.

Newton's method: $\min f(x)$. Define $g_t = g(x_t) = \frac{\partial f(x_t)}{\partial x}$ and $H_t = \frac{\partial^2 f(x_t)}{\partial x \partial x'}$

Then Newton's method implements the iteration: $x_{t+1} = x_t - H_t^{-1}g_t$

6 Measure Theory*

We shall assume some familiarity with basic concepts on measure theory which the modern probability theory is based on.

- σ field. This is a collection \mathfrak{F} of subsets of a set Ω that satisfies
 1. $\Omega \in \mathfrak{F}$
 2. $F \in \mathfrak{F} \Rightarrow F^c \in \mathfrak{F}$
 3. $F_n \in \mathfrak{F}, \forall n \Rightarrow \bigcup_1^\infty F_n \in \mathfrak{F}$.
- (Ω, \mathfrak{F}) is called a measurable space.
- A σ -finite measure m on (Ω, \mathfrak{F}) is a nonnegative σ -additive set function $m : \mathfrak{F} \rightarrow R_+ = [0, \infty)$ satisfying
 1. $m(\Phi) = 0$
 2. $m(\bigcup_1^\infty F_n) = \sum_1^\infty m(F_n)$ for any sequence $\{F_n\}$ of pairwise disjoint member of \mathfrak{F} (ie $F_n \cap F_m = \Phi, \forall n \neq m$).

- A measure space $(\Omega, \mathfrak{F}, m)$ is a triplet where (Ω, \mathfrak{F}) is a measurable space and m is a σ -finite measure on (Ω, \mathfrak{F}) .
- A probability space $(\Omega, \mathfrak{F}, P)$ is a normalized measurable space in which $P(\Omega) = 1$. P is then called a probability measure on (Ω, \mathfrak{F}) .
- $\mathfrak{B} = \mathfrak{B}(R)$ is the Borel σ -field of R , ie, the σ field generated by all open subsets of R .
- Let $(\Omega, \mathfrak{F}, P)$ be a probability space. Then a function $f : \Omega \rightarrow R$ is measurable (or \mathfrak{F} measurable) if

$$f^{-1}B \in \mathfrak{F}, \forall B \in \mathfrak{B}.$$

Here $f^{-1}B$ is the preimage of B under f and the requirement is that f takes measurable sets (in \mathfrak{F}) into measurable sets in \mathfrak{B} .

- If $X : \Omega \rightarrow R$ is \mathfrak{F} -measurable (ie $\{\omega | X(\omega) \leq r\} \in \mathfrak{F}$ for all $r \in R$) then we call X a real valued random variable on Ω .

- Examples:

1. Indicator random variable. Let $A \in \mathfrak{F}$. Define:

$$1_A : \Omega \rightarrow R \text{ by } 1_A(w) = 1 \text{ if } w \in A \text{ and } 0 \text{ otherwise}$$

2. Simple random variable. Let $\{A_i\}_1^n \in \mathfrak{F}$ be disjoint subsets of Ω ($A_i \cap A_j = \Phi$ if $i \neq j$).

Define

$$f : \Omega \rightarrow R$$

by

$$f(w) = \sum_1^n a_i 1_{A_i}(w), \quad a_i > 0$$

Then f is a simple random variable.

- Integral of a simple function. If $f = \sum_1^n a_i$ is a simple function, we define the integral of f with respect to P as

$$\int f dP = \sum_1^n a_i P(A_i).$$

- Suppose $X : \Omega \rightarrow \bar{R}_+$ ($\equiv R_+ \cup \{\infty\}$). Then $X(w) \geq 0$ is a random variable on $(\Omega, \mathfrak{F}, P)$ iff $X(w)$ is a pointwise limit of an increasing sequence of simple random variables X_n . The idea of the proof is to use the following construction:

$$X_n(w) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbf{1} \left(\frac{k-1}{2^n} \leq X(w) \leq \frac{k}{2^n} \right) + n \mathbf{1}(X(w) \geq n)$$

- Integral of a nonnegative function. If F is \mathfrak{F} -measurable and $f \geq 0$, we define

$$\int f dP = \lim_{n \rightarrow \infty} \int f_n dP$$

where $\{f_n\}$ is a sequence of sample random variables such that $f_n \rightarrow f$.
 f is integrable if $\int f dP < \infty$.

- Integral of an \mathfrak{F} -measurable function. If $f : \Omega \rightarrow R$ is \mathfrak{F} -measurable, set $f^+ = \max(f, 0)$ and $f^- = \max(-f, 0)$ and write

$$f = f^+ - f^-$$

Then $f^+ \geq 0, f^- \geq 0$ and we have sequences of sample random variables

$$f_n^+ \rightarrow f^+, f_n^- \rightarrow f^-$$

and we define

$$f_n = f_n^+ - f_n^-$$

and

$$\int f dP = \lim_{n \rightarrow \infty} \int f_n dP.$$

We say f is measurable if

$$\int f^+ dP < \infty, \int f^- dP < \infty.$$

- Let $L_0(P) = L_0(\Omega, \mathfrak{F}, P)$ denote the space of all real valued random variables on $(\Omega, \mathfrak{F}, P)$, ie,

$$L_0(P) = \{X | X : \Omega \rightarrow R, X \text{ is measurable}\}$$

- Let

$$L_r(P) = \{X | X : \Omega \rightarrow R, \int |X|^r dP < \infty\}$$

When $r \geq 1$, this space is a real Banach space (complete normed matrix space) with norm

$$\| X \|_r = (\int |X|^r dP)^{1/r} = (E|X|^r)^{1/r}$$

- $L_2(P)$ is a Hilbert space with inner product given by

$$(X, Y) = \int XY dP, X, Y \in L_2(P)$$

and by Schwarz's inequality this is bounded by

$$|(X, Y)| \leq \| X \|_2 \| Y \|_2$$

The geometric structure of $L_2(P)$ is useful in defining projections.

- Coordinate presentation of random sequence. It is helpful to identify an abstract probability space $(\Omega, \mathfrak{F}, P)$ and a random sequence $\{X_n\}_{-\infty}^{\infty}$ defined on it by their coordinate representations

1. Define $h : \Omega \rightarrow R_{\infty} (\equiv \times_{-\infty}^{\infty} R)$ by

$$h(\omega) = (\cdots, X_{-1}(\omega), X_0(\omega), X_1(\omega), \cdots) = (\cdots, x_{-1}, x_0, x_1, \cdots) = x$$

2. $\mathfrak{B}_{\infty} = \mathfrak{B}(R_{\infty})$, Borel σ -field on R_{∞} .

3. $X_n : R_{\infty} \rightarrow R$ coordinate functions defined by $X_n(x) = x_n$.
Hence, $\{x | X_n(x) \leq a\} = \{x | \cdots, X_{n-1} < \infty, X_n \leq a, X_{n+1} \leq \infty, \cdots\} \in \mathfrak{B}_{\infty}$

- Strict stationarity: $\{X_n\}_1^\infty$ is strictly stationary if finite dimensional distributions are translation invariant, ie,

$$(X_{t_1+h}, \dots, X_{t_p+h}) = (X_{t_1}, \dots, X_{t_p}), \forall h, p, t_1, \dots, t_p$$

- Temporal displacements and shift operator. $(R_\infty, \mathfrak{B}_\infty, P)$ with typical element of $R_\infty : x = (\dots, x_{-1}, x_0, x_1, \dots)$. Backshift operator S is defined as: $Sx = (\dots, x_0, x_1, x_2, \dots)$.
- Denote U_S the induced operator. Observe that if $\{X_n\}$ is a sequence on $(R_\infty, \mathfrak{B}_\infty, P)$ then

$$\begin{aligned} X_1(x) &= x_1 \\ X_2(x) &= X_1(Sx) = x_2 \\ X_3(x) &= X_1(S^2x) = x_3 \\ &\vdots \\ X_{n+1}(x) &= X_1(S^n x) = x_n \end{aligned}$$

inducing

$$U_S : L_0(R_\infty, \mathfrak{B}_\infty, P) \rightarrow L_0(R_\infty, \mathfrak{B}_\infty, P)$$

defined by

$$U_S X(x) = X(Sx)$$

and

$$U_S X_n = X_{n+1}, X_n = U_S^{n-1} X_1$$

- Stationary sequences $\{X_n\}$ give translation invariance

$$P(E) = P(S^{-h}E), \forall h,$$

ie P is preserved under the action of S

- Definition. $S : \Omega \rightarrow \Omega$ is a measure preserving if (1) S is measurable; (2) $P(S^{-1}B) = P(B), \forall B \in \mathfrak{F}$.
- Theorem: If (1) $X = \{X_n\}_{-\infty}^{\infty}$ is strictly stationary; (2) $\varphi : R_{\infty} \rightarrow R$ is measurable (ie $\varphi^{-1}B \in R_{\infty}, \forall B \in \mathfrak{B}$); (3) $Y_n = \varphi(\dots, x_{n-1}, x_n, x_{n+1}, \dots)$. Then $\{Y_n\}$ is strictly stationary.
- Can we generalize Kolmogorov LLN to temporally dependent sequences?
A counterexample: $X_t = u_t + Z$, where U_t is an iid uniform $[0, 1]$, Z is $N(0,1)$ and independent of u_t . Obviously $E(X_t) = 1/2$, but $\bar{X} = \bar{u} + Z \xrightarrow{a.s.} 1/2 + Z$. Why LLN breaks down? Too much dependence in the sequence of X_t .

- Ergodicity. When do temporal averages such as

$$\frac{1}{n} \sum_1^n X_j, \frac{1}{n} \int X(s) ds$$

converges to spatial averages $E(X)$, ie, when does the average of a physical system over time is the same as the average of infinitely many identical systems at one point in time?

- Several concepts: Given $(\Omega, \mathfrak{F}, P)$, $S : \Omega \rightarrow \Omega$ is a measure preserving map. (1) An event F is invariant if $F = S^{-1}F$; (2) S is ergodic if all for invariant events F , $P(F) = 0, 1$; (3) Strictly stationary process $\{X_t\}$ ($X_t = U_s^{t-1}X_1$) is ergodic if S is ergodic.

- Remarks

1. Absence of ergodicity means that there exist invariant events F for which $0 < P(F) < 1$.
2. Hence it is impossible to fully sample Ω if we start off in F .
3. S does not properly mix the points of Ω .

- Theorem. If (1) $X \in L_1(\Omega, \mathfrak{F}, P)$; (2) $S : \Omega \rightarrow \Omega$ is measure preserving and ergodic. Then

$$P \left(\lim_{n \rightarrow \infty} \frac{1}{n} S_n = E(x) \right) = 1,$$

ie

$$\bar{X} \xrightarrow{a.s.} E(X).$$

- Theorem (necessary and sufficient condition for ergodicity): Let $(\Omega, \mathfrak{F}, P)$ be a probability space and $S : \Omega \rightarrow \Omega$ be a measure preserving map on Ω . S is ergodic iff

$$\frac{1}{n} \sum_{k=0}^{n-1} P(F \cap S^{-k}G) \rightarrow P(F)P(G), \forall F, G \in \mathfrak{F}.$$

- Theorem: If (1) $X = \{X_n\}_{-\infty}^{\infty}$ is strictly stationary and ergodic; (2) $\varphi : R_{\infty} \rightarrow R$ is measurable (ie $\varphi^{-1}B \in R_{\infty}, \forall B \in \mathfrak{B}$); (3) $Y_n = \varphi(\cdots, x_{n-1}, x_n, x_{n+1}, \cdots)$. Then $\{Y_n\}$ is strictly stationary and ergodic.