STA347 Notes

Ian Zhang

July 29, 2024

Contents

1 Expectation **3**
- 1.1 Average Operator 3
- 1.2 Definition of Expectation 3
- 1.3 Examples of Expectation 4
- 1.4 Moments 6
- 1.5 Sample Surveys 6
- 1.6 Least Squares Estimation 7

2 Probability **7**
- 2.1 Indicator Functions 7
- 2.2 Probabilities 9
- 2.3 Inequalities 9
- 2.4 Product Moment Matrices 11
  - 2.4.1 Cauchy-Schwarz Inequality 12
- 2.5 Principle of Inclusion-Exclusion 12
- 2.6 Independence 12
  - 2.6.1 Independence of Events 14
- 2.7 Generating Functions 14
  - 2.7.1 Exponential Distribution 15
  - 2.7.2 Gamma Distribution 15

3 Conditioning **16**
- 3.1 Conditional Expectation 16
- 3.2 Conditional Probability 19
- 3.3 Independence from a Conditional Perspective 20

4 Continuous Random Variables and Their Transformations **21**
4.1 Distributions with a Density 21
4.1.1 Transformations 21
4.2 Conditional Densities 22
4.3 Order Statistics 23
4.3.1 Distribution of Order Statistics 23

5 Basic Limit Theorems **24**
5.1 Convergence in Probability 24
5.1.1 Weak Law of Large Numbers 25
5.2 Convergence in Distribution 26
5.2.1 Normal Random Variables 27
5.2.2 Central Limit Theorem 27

1 Expectation

Let Ω be the sample space and $\omega \in \Omega$ be points in the sample space. A random variable is a function $X : \Omega \to \mathbb{R}$, so we consider $X(\omega)$ for $\omega \in \Omega$.

1.1 Average Operator

Consider a finite sample space Ω that consists of n_i ω_i's for $i = 1, \ldots, k$ and let $n = n_1 + \ldots + n_k$. For any random variable X, define

$$A(X) := \frac{1}{n} \sum_{\omega \in \Omega} X(\omega) = \sum_{i=1}^{k} \frac{n_i}{n} X(\omega_i) = \sum_{i=1}^{k} p_i X(\omega_i)$$

where $p_i = \frac{n_i}{n}$ so $\sum_{i=1}^{k} p_i = 1$, $p_i \geq 0$ (p_i represents the *proportion* of elements in Ω that are ω_i). The properties of A are

1. If $X \geq 0$, then $A(X) \geq 0$
2. If X, Y are random variables, then $A(c_1X + c_2Y) = c_1A(X) + c_2A(Y)$ where c_1, c_2 are constants
3. $A(1) = 1$

Proof. To show 1., suppose $X(\omega) \geq 0$ for all $\omega \in \Omega$. Then since each $p_i \geq 0$, it follows that $p_iX(\omega_i) \geq 0$, thus $A(X) \geq 0$ by transitivity.
To show 2., by definition of A,

$$A(c_1X + c_2Y) = \sum_{i=1}^{k} p_i[c_1X(\omega_i) + c_2Y(\omega_i)] = c_1 \sum_{i=1}^{k} p_iX(\omega_i) + c_2 \sum_{i=1}^{n} p_iY(\omega_i) = c_1A(X) + c_2A(Y)$$

To show 3., note that since $1(\omega) = 1$ for all $\omega \in \Omega$, then

$$A(1) = \sum_{i=1}^{k} p_i = 1$$

by assumption of the p_i's. ■

1.2 Definition of Expectation

An operator E is an expectation operator if it satisfies the following axioms:

1. If $A \geq 0$, then $E(X) \geq 0$
2. If X, Y are random variables, then $E(c_1X + c_2Y) = c_1E(X) + c_2E(Y)$ where c_1, c_2 are constants
3. $E(1) = 1$

4. For $X_1, X_2, \ldots \geq 0$, if $X_n \uparrow X$, then $E(X_n) \uparrow X$.
   - This properties does *not* imply that $X_i \to X \implies E(X_i) \to E(X)$; we must have $X_i \uparrow X$ to confidently assert any sort of convergence of expectation

Properties:

(a) $E(c_1X_1 + \cdots + c_nX_n) = c_1E(X_1) + \cdots + c_nE(X_n)$

(b) If $X \leq Y$, then $E(X) \leq E(Y)$

(c) $|E(X)| \leq E(|X|)$

(d) (Fatou's Lemma) If $X_n(\omega) \geq 0$ and $X_n(\omega) \to X(\omega)$, then $\liminf_n E(X_n) \geq E(X)$

Definition. Let $(a_i)_i$ be a sequence of real numbers and define the sequence $(b_i)_i$ where

$$b_i := \inf_{k \geq i} a_i$$

Then

$$\liminf_i a_i = \lim_{i \to \infty} b_i$$

Similarly,

$$\limsup_i a_i = -\liminf_i(-a_i) = \lim_{i \to \infty} \left(\sup_{k \geq i} a_i \right)$$

Proposition. A sequence $(a_i)_i$ converges to a iff

$$\liminf_i a_i = \limsup_i a_i = a$$

Theorem (Dominated Convergence)**.** If $X_n(\omega) \to X(\omega)$ and $|X_n(\omega)| \leq Y(\omega)$ for all $n \in \mathbb{N}$, $\omega \in \Omega$, and $E(Y) < \infty$, then $E(X_n) \to E(X)$.

- Y is called a dominator of X_n

1.3 Examples of Expectation

Theorem. The sample space Ω is discrete with elements $\{\omega_1, \ldots, \omega_k\}$ iff the expectation operator takes the form

$$E(X) = \sum_{i=1}^{k} p_i X(\omega_i)$$

where $p_i \geq 0$ for all i and $\sum_{i=1}^{n} p_i = 1$.

- To show a sample space Ω is discrete, we can show that there exists a discrete subset of Ω with probability 1 (we can say this subset is essentially the entire sample space)

Proof. To show sufficiency, note that

$$X(\omega) = \sum_{i=1}^{k} I(\{\omega = \omega_i\})X(\omega_i)$$

Take

$$\begin{aligned} E(X) &= E\left(\sum_{i=1}^{k} I(\{\omega = \omega_i\})X(\omega_i)\right) \\ &= \sum_{i=1}^{k} E(I(\{\omega = \omega_i\})X(\omega_i) \\ &= \sum_{i=1}^{k} P(\omega_i)X(\omega_i) \end{aligned}$$

where we take $p_i = P(\omega_i)$. Setting $X = 1$, this shows $\sum_{i=1}^{k} p_i = 1$.
To show necessity, take $X = I\{\omega = \omega_1\})$, thus $E(X) = P(\omega_1)$ and $\sum_{i=1}^{k} p_i = p_1$, so $P(\omega_1) = p_1$. Similarly, for all i, $p_i = P(\omega_i)$. Thus since the $\{\omega_i\}$ are discrete,

$$\begin{aligned} &P\left(\bigcup_{i=1}^{k}\{\omega_i\}\right) = 1 \\ \implies &\bigcup_{i=1}^{k}\{\omega_i\} \text{ is essentially the entire sample space} \\ \implies &\Omega \text{ is essentially a discrete space with realizations } \omega_1, \ldots, \omega_k \end{aligned}$$

■

Definition (Continuous Random Variables)**.** Let $\Omega = \mathbb{R}$. A random variable X is continuous if there exists a continuous $f \geq 0$ with

$$\int_{\mathbb{R}} f(x)\,dx = 1$$

such that

$$E(X) = \int_{-\infty}^{\infty} X(\omega)f(\omega)\,d\omega$$

Suppose $X = I(A)$ for some subset $A \subseteq \Omega$. Then

$$P(A) = \int_A f(\omega)\,d\omega$$

Note that the above equations are equivalent to

$$E[H(X)] = \int_{\mathbb{R}} H(x)f(x)\,dx$$

and

$$P(X \in A) = \int_A f(x)\,dx$$

1.4 Moments

Definition. If X is a random variable, define its jth moment to be

$$\mu_j = E(X^j)$$

1.5 Sample Surveys

Set up N individuals $\omega_1, \dots, \omega_N$ and select a sample

$$(\xi_1, \dots, \xi_n)$$

Let $Z_i = X(\xi_i)$ for all i and define

$$\bar{Z} = \frac{1}{n}(Z_1 + \cdots + Z_n)$$

Denote $x_k = X(\omega_k)$ for $k \in \{1, \dots, N\}$. Since each Z_i has equal probability of taking on any x_k value, then

$$E(Z_i) = \frac{1}{N}\sum_{i=1}^{N} x_i =: \bar{X}$$

By linearity,

$$E(\bar{Z}) = \frac{1}{n}E\left(\sum_{i=1}^{n} Z_i\right) = \bar{X}$$

By symmetry, it holds that

$$E(Z_i^2) = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

Thus

$$\mathrm{Var}(Z_i) = E(Z_i^2) - \bar{X}^2 =: V(X)$$

Theorem. If sampling is without replacement, then

$$E(\bar{Z}) = \bar{X}$$

and

$$\mathrm{Var}(\bar{Z}) = \frac{1}{n}\frac{N-n}{N-1}V(X)$$

If sampling is with replacement, then

$$E(\bar{Z}) = \bar{X}$$

and

$$\mathrm{Var}(\bar{Z}) = \frac{1}{n}V(X)$$

1.6 Least Squares Estimation

Given a response variable X and predictor variables $Y_1, \ldots, Y_m$, we want to predict X using the information we have (Y_i) by minimizing

$$E[(X - a_0 - a_1Y_1 + \cdots + a_mY_m)^2]$$

Represent the Y_i as a vector

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_m \end{bmatrix}$$

Define the covariance matrix of Y to be a symmetric matrix

$$\text{Cov}(Y) = \begin{bmatrix} \text{Var}(Y_1) & \text{Cov}(Y_1, Y_2) & \cdots & \text{Cov}(Y_1, Y_m) \\ \text{Cov}(Y_2, Y_1) & \text{Var}(Y_2) & \cdots & \text{Cov}(Y_2, Y_m) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Y_1, Y_m) & \text{Cov}(Y_2, Y_m) & \cdots & \text{Var}(Y_m) \end{bmatrix}$$

and the cross-covariance matrix of Y and X to be

$$\text{Cov}(Y, X) = \begin{bmatrix} \text{Cov}(Y_1, X) \\ \vdots \\ \text{Cov}(Y_m, X) \end{bmatrix}$$

Theorem. The best linear predictor of X is

$$\hat{X} = a_0 + a_1Y_1 + \cdots + a_mY_m$$

where $a^T = \begin{bmatrix} a_1 & \cdots & a_m \end{bmatrix}$ satisfies

$$\text{Cov}(Y)a = \text{Cov}(Y, X)$$

and

$$a_0 = E(X) - \sum_{j=1}^{m} a_jE(Y_j)$$

2 Probability

2.1 Indicator Functions

For simplicity, denote $I(A) = I_A(\omega)$ for all $\omega \in \Omega$.
<u>**Properties:**</u>

1. $I(A^c) = 1 - I(A)$

2. If $A \subseteq B$, then $I(A) \leq I(B)$
3. $I(A \cup B) = \max\{I(A), I(B)\}$
4. $I(A \cap B) = \min\{I(A), I(B)\}$
5. If $A_1 \subseteq A_2 \subseteq \cdots$, then $I\left(\bigcup_{i=1}^{\infty} A_i\right) = \sup_{i\geq 1} I(A_i) = \lim_{i\to\infty} I(A_i)$

Proof. If

$$I(A) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

then

$$1 - I(A) = \begin{cases} 1 & \omega \notin A \\ 0 & \omega \in A \end{cases} = I(A^c)$$

Suppose $A \subseteq B$. Consider the following 3 cases:

1. $\omega \in A \implies \omega \in B \implies I_A(\omega) = 1 = I_B(\omega)$
2. $\omega \in B \setminus A \implies I_A(\omega) = 0 < 1 = I_B(\omega)$
3. $\omega \notin B \implies \omega \notin A \implies I_A(\omega) = 0 = I_B(\omega)$

which shows $I(A) \leq I(B)$.
Consider $A \cup B$ and the following 4 cases:

1. $\omega \in A \cup B \setminus A \implies I_{A\cup B} = 1 = \max\{0, 1\} = \max\{I_A(\omega), I_B(\omega)\}$
2. $\omega \in A \cup B \setminus B \implies I_{A\cup B} = 1 = \max\{1, 0\} = \max\{I_A(\omega), I_B(\omega)\}$
3. $\omega \in A \cap B \implies \omega \in A \cup B \implies I_{A\cup B}(\omega) = 1 = \max\{1, 1\} = \max\{I_A(\omega), I_B(\omega\}$
4. $\omega \notin A \cup B \implies x \notin A, x \notin B \implies I_{A\cup B} = 0 = \max\{0, 0\} = \max\{I_A(\omega), I_B(\omega)\}$

Consider $A \cap B$ and the following cases:

1. $\omega \in A \cap B \implies I_{A\cap B}(\omega) = 1 = \min\{1, 1\} = \min\{I_A(\omega), I_B(\omega\})$
2. $\omega \in A \setminus A \cap B \implies \omega \in A, \omega \notin B \implies \omega \notin A \cap B \implies I_{A\cap B}(\omega) = 0 = \min\{1, 0\} = \min\{I_A(\omega), I_B(\omega)\}$
3. $\omega \in B \setminus A \cap B \implies \omega \notin A, \omega \in B \implies \omega \notin A \cap B \implies I_{A\cap B}(\omega) = 0 = \min\{0, 1\} = \min\{I_A(\omega), I_B(\omega)\}$
4. $\omega \notin A, \omega \notin B \implies \omega \notin A \cap B \implies I_{A\cap B}(\omega) = 0 = \min\{0, 0\} = \min\{I_A(\omega), I_B(\omega)\}$

Suppose $A_1 \subseteq A_2 \subseteq \cdots$. Consider the following cases:

1. If $\omega \in \bigcup_{i=1}^{\infty} A_i$, then there exists $k \in \mathbb{N}$ such that $\omega \in A_k$, so $\omega \in A_j$ for all $j \geq k$, thus $I_{A_j}(\omega) = 1$ for all $j \geq k \geq 1$, so $\sup_{i \geq 1} I(A_i) = 1$. Moreover, this also shows that
$$\lim_{i \to \infty} I(A_i) = 1$$
Since $\omega \in \bigcup_{i=1}^{\infty} A_i$, then $I\left(\bigcup_{i=1}^{\infty} A_i\right) = 1$.

2. If $\omega \notin \bigcup_{i=1}^{\infty} A_i$, then for all $i \in \mathbb{N}$, $\omega \notin A_i$, thus $I_{A_i}(\omega) = 0$ for all i. This implies $\sup_{i \geq 1} I_{A_i}(\omega) = 0$ and
$$\lim_{i \to \infty} I_{A_i}(\omega) = 0$$
Since $\omega \notin \bigcup_{i=1}^{\infty} A_i$, then $I\left(\bigcup_{i=1}^{\infty} A_i\right) = 0$, which proves our claim.

■

2.2 Probabilities

Definition. Let $A \subseteq \Omega$. Let I_A be the indicator function on A. The probability of A is

$$P(A) = E(I_A)$$

Properties:

1. $0 \leq P(A) \leq I$
2. $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
3. $P(\Omega) = 1$
4. If $A_1 \subseteq A_2 \subseteq \cdots$, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{i \to \infty} P(A_i)$

We prove these properties using the properties of indicator functions.

2.3 Inequalities

Proposition. Suppose X is a nonnegative random variable. Then for all $a > 0$, we have

$$I(\{X(\omega) > a\}) \leq \frac{X(\omega)}{a}$$

for all $\omega \in \Omega$.

Proof. Suppose for some ω that $X(\omega) > a$, thus

$$I(\{X(\omega) > a\}) = 1 < \frac{X(\omega)}{a}$$

Suppose for some ω that $X(\omega) \leq a$, since X is nonnegative and a is positive, then

$$\frac{X(\omega)}{a} \geq 0 = I(\{X(\omega) > a\})$$

as required. ■

From this identity, we can deduce Markov's Inequality:

Corollary (Markov's Inequality)**.** For any nonnegative random variable X and $a > 0$,

$$P(X > a) \leq \frac{E(X)}{a}$$

If we take $X = |Y - E(Y)|$ for some random variable Y, then we have Chebyshev's Inequality:

Corollary (Chebyshev's Inequality)**.** If Y is a random variable and $a > 0$,

$$P(|Y - E(Y)| > a) \leq \frac{\text{Var}(Y)}{a^2}$$

Proof. By definition of absolute value, $|Y - E(Y)| \geq 0$, thus by Markov's Inequality,

$$P(|Y - E(Y)| > a) = P((Y - E(Y))^2 > a^2) \leq \frac{E[(Y - E(Y))^2]}{a^2} = \frac{\text{Var}(Y)}{a^2}$$

■

Proposition. If $X \geq 0$, then $E(X) = \int_0^\infty P(X > t)\, dt$

Proof. Rewrite

$$X = \int_0^X 1\, dt = \int_0^\infty I(t < X)\, dt$$

By the infinite sum nature of the Riemann integral,

$$\begin{aligned} E\left(\int_0^\infty I(t < X)\, dt\right) &= \int_0^\infty E(I(t < X))\, dt \\ &= \int_0^\infty P(t < X)\, dt \end{aligned}$$

as required. ■

Theorem. If $X \geq 0$, then $E(X) = 0$ iff $X = 0$ almost surely (i.e., $P(X = 0) = 1$).

Proof. Suppose $E(X) = 0$. Define events $A_k = \left\{X > \frac{1}{k}\right\}$, which form an increasing sequence of events. As $k \to \infty$, $A_k \to \{X > 0\} = \bigcup_{k=1}^\infty A_k$. By property of probability, $P(A_k) \to P(X > 0)$. On the other hand, by Markov's Inequality, since X is nonnegative and $\frac{1}{k} > 0$,

$$0 \leq P(A_k) = P\left(X > \frac{1}{k}\right) \leq \frac{E(X)}{\frac{1}{k}} = 0$$

thus $P(A_k) = 0$ for all k. By uniquness of the limit, $P(A_k) \to 0$ implies $P(X > t) = 0$, so $P(X = 0) = 1$, as required.
Suppose $P(X = 0) = 1$. This implies $P(X > 0) = 0$, thus

$$E(X) = \int_0^\infty 0\, dt = 0$$

as required. ■

Corollary. If X is a random variable, then $\text{Var}(X) = 0$ iff $X = \mu$ for almost surely where μ is constant.

Proof. Suppose $\text{Var}(X) = 0$. By definition,

$$E[(X - E(X))^2] = 0$$

which implies $(X - E(X))^2 = 0$ almost surely since $(X - E(X))^2 \geq 0$. This implies $X = E(X)$ almost surely, and taking $\mu = E(X)$ proves sufficiency.
Suppose $X = \mu$ almost surely. Then $|X - \mu| = 0$ almost surely, thus $E(|X - \mu|) = 0$, which implies $E(X) = \mu$. This implies $|X - E(X)|^2 = 0$ almost surely, so $\text{Var}(X) = E[|X - E(X)|^2] = 0$. ■

2.4 Product Moment Matrices

Definition. If $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}^T$ is a random vector, then $U = E(XX^T)$ is the product moment matrix.

- By definition, if $Y = X - E(X)$, then the product moment matrix of Y is the covariance matrix of X.

Theorem. A product moment matrix U is symmetric and positive semidefinite. It is singular iff $c^T X = 0$ almost surely for some constant vector c.

Proof. Since XX^T is symmetric, then $E(XX^T)$ is also symmetric. For any vector a,

$$a^T U a = a^T E(XX^T) a = E(a^T XX^T a) = E[(a^T X)^2] \geq 0$$

since $(a^T X)^2 \geq 0$, thus U is positive semidefinite by definition.
To show the rest of the claim,

$$\begin{aligned} U \text{ is singular} &\iff \det(U) = 0 \\ &\iff 0 \text{ is an eigenvalue of } U && \text{det is the product of eigenvalues} \\ &\iff c^T U c = 0 \\ &\iff E(c^T XX^T c) = 0 \end{aligned}$$

$$\iff E[(c^T X)^2] = 0$$
$$\iff c^T X = 0 \quad a.s. \qquad\qquad E[(c^T X)^2] = 0 \text{ implies } (c^T X)^2 = 0 \ a.s.$$

■

2.4.1 Cauchy-Schwarz Inequality

If X_1, X_2 are random variables, then

$$[E(X_1 X_2)]^2 \leq (E(X_1^2))(E(X_2^2))$$

with equality holding iff $c_1 X_1 + c_2 X_2 = 0$ almost surely for some constants c_1, c_2 satisfying $c_1^2 + c_2^2 \neq 0$.

Proof. Consider the random vector $X^T = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$ and its product moment matrix

$$U = E\left(\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}\begin{bmatrix} X_1 & X_2 \end{bmatrix}\right) = \begin{bmatrix} E(X_1^2) & E(X_1 X_2) \\ E(X_1 X_2) & E(X_2^2) \end{bmatrix}$$

Since U is positive semidefinite, $\det(U) \geq 0$, thus $E(X_1^2)E(X_2^2) - (E(X_1 X_2))^2 \geq 0$, which shows the inequality.
Note that equality holds iff U is singular iff there exists c_1, c_2 such that $c_1^2 + c_2^2 \neq 0$ and $c_1 X_1 + c_2 X_2 = 0$ almost surely. ■

2.5 Principle of Inclusion-Exclusion

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i) + \sum_{i_1 < i_2} P(A_{i_1} \cap A_{i_2}) + \cdots + (-1)^{r+1} \sum_{i_1 < i_2 < \cdots < i_r} P\left(\bigcap_{j=1}^r A_{i_j}\right) + \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^n A_i\right)$$

2.6 Independence

Suppose we have a spatial region with M cells and N molecules. Let ξ_i be the position of the ith molecule. There are M^N elements in the sample space of possible positions for the N molecules. Suppose that all the molecules are distributed uniformly and define

$$E[X(\omega)] = \frac{1}{M^N} \sum_{a_1=1}^M \cdots \sum_{a_N=1}^M X(a_1, \ldots, a_n) \tag{1}$$

where $\omega^T = \begin{bmatrix} a_1 & \cdots & a_N \end{bmatrix}$ is a possible positioning in the sample space.

Theorem. (1) implies that the $\xi_1, \ldots, \xi_N$ are uniformly distributed over $\{1, \ldots, M\}$ and

$$E\left[\prod_{k=1}^{N} H_k(\xi_k)\right] = \prod_{k=1}^{N} E[H_k(\xi_k)]$$

for all H_k, $k \in \{1, \ldots, N\}$.

Proof. Let $X = I(\xi_i = k)$ for all $i \in \{1, \ldots, N\}$ and $k \in \{1, \ldots, M\}$. Since $\xi_i(\omega) = \omega_i$, then $X = I(\omega_i = k) = k$ where $\omega^T = \begin{bmatrix}\omega_1 & \cdots & \omega_N\end{bmatrix}$ By (1),

$$E(X) = \frac{1}{M^N} \sum_{a_1=1}^{M} \cdots \sum_{a_N=1}^{M} X(a_1, \ldots, a_N)$$

Since $X(\omega) = 0$ unless $\omega_i = k$ and there are M^{n-1} possible $\omega \in \Omega$ such $\omega_i = k$, then

$$P(w_i = k) = \frac{M^{N-1}}{M^N} = \frac{1}{M}$$

which shows that the ξ_i are uniformly distributed.
To show the rest of the claim, notice

$$\begin{aligned} E\left[\prod_{k=1}^{N} H_k(\xi_k)\right] &= \frac{1}{M^N} \sum_{a_1=1}^{M} \cdots \sum_{a_N=1}^{M} \left(\prod_{k=1}^{N} H_k(a_k)\right) \\ &= \frac{1}{M^N} \left(\sum_{a_1=1}^{M} H_1(a_1)\right) \cdots \left(\sum_{a_N=1}^{M} H_N(a_N)\right) \\ &= \frac{1}{M^N} \prod_{k=1}^{N} \sum_{a_k=1}^{M} H_k(a_k) \end{aligned}$$

but since the molecules are uniformly distributed, then

$$E[H_k(\xi_k)] = \frac{1}{M} \sum_{i=1}^{M} H_k(\xi_i)$$

thus

$$\prod_{k=1}^{N} E[H_k(\xi_k)] = \prod_{k=1}^{N} \frac{1}{M} \sum_{i=1}^{M} H_k(\xi_i) = \frac{1}{M^N} \prod_{k=1}^{N} \sum_{a_k=1}^{M} H_k(a_k)$$

as required. ∎

Definition. Random variables $X_1, \ldots, X_p$ are independent if

$$E\left[\prod_{i=1}^{p} H_i(X_i)\right] = \prod_{i=1}^{p} E[H_i(x_i)]$$

for all functions $H_1, \ldots H_p$.

Proposition. $X_1, \ldots, X_p$ are independent iff $P(X_1 \in A_1, \ldots, X_p \in A_p) = \prod_{i=1}^{p} P(X_i \in A_i)$ for all $A_i \subseteq \Omega$ and $i = 1, \ldots, p$.

Proposition. Define cdf $F(x_1, \ldots, x_p)$ as the joint cdf of $X_1, \ldots, X_p$. Then $X_1, \ldots, X_p$ are independent iff

$$F(x_1, \ldots, x_p) = \prod_{i=1}^{p} F(x_i) \tag{2}$$

Note: pmf/pdf's are only defined for certain classes of random variables but cdfs are defined for all.

Corollary. If X_1 and X_2 are discrete and take integer values, then $X_1 \perp\!\!\!\perp X_2$ iff $P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$ for all $x_1, x_2 \in \mathbb{Z}$.

2.6.1 Independence of Events

Definition. Events $A_1, \ldots$ are independent if the indicator random variables $I(A_1), \ldots$ are independent.

Proposition. $A_1, A_2, \ldots$ are independent iff

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j})$$

Note: Pairwise independence does not imply joint independence.

2.7 Generating Functions

Definition. If X is a random variable, define its probability generating function as

$$\Pi(z) = E(z^X), z > 0$$

and its moment generating function as

$$M_X(z) = E(e^{zX}), z \in \mathbb{R}$$

Theorem. If X and Y are independent, then

$$\Pi_{X+Y}(z) = \Pi_X(z)\Pi_Y(z)$$
$$M_{X+Y}(z) = M_X(z)M_Y(z)$$

Proof. Follows by definition of independence. ■

Theorem. If X and Y are random variables and

$$\Pi_X(z) = \Pi_Y(z) < \infty \quad \forall z \in [1-\delta, 1+\delta] \text{ for some } \delta > 0$$

or

$$M_X(z) = M_Y(z) < \infty \quad \forall z \in [-\delta, \delta] \text{ for some } \delta > 0$$

then X and Y are identically distributed.

Theorem. If $X \sim \text{Poisson}(\lambda)$, $Y \sim \text{Poisson}(\mu)$ and $X \perp\!\!\!\perp Y$, then

$$X + Y \sim \text{Poisson}(\lambda + \mu)$$

Proof. By computation, the mgf of Poisson(α) is

$$M(z) = \sum_{i=0}^{\infty} P(X = i)e^{zi} = \exp(-\alpha) \sum_{i=0}^{\infty} \frac{(\alpha \exp(z))^i}{i!} = \exp(\alpha(\exp(z) - 1))$$

Since X and Y are independent, then

$$\begin{aligned} M_{X+Y}(z) &= M_X(z)M_Y(z) \\ &= \exp(\lambda(\exp(z) - 1)) \exp(\mu(\exp(z) - 1)) \\ &= \exp((\lambda + \mu)(\exp(z) - 1)) \end{aligned}$$

which is the mgf of a Poisson($\lambda + \mu$) distribution. ■

Theorem. If $M_X(z) < \infty$ for $z \in [-\delta, \delta]$ for some $\delta > 0$, then

$$E(X^k) = M_X^{(k)}(0)$$

2.7.1 Exponential Distribution

Definition. A random variable X is Exponential with parameter λ if its cdf is

$$F(x) = 1 - \exp(-\lambda x), x \geq 0$$

2.7.2 Gamma Distribution

The **Gamma function** is given by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} \, dx$$

for $\alpha > 0$. Its properties include

1. $\Gamma(\alpha + 1) = \Gamma(\alpha)$
2. $\Gamma(n) = (n - 1)!$ for all $n \in \mathbb{N}$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

Definition. A random variable X has Gamma distribution with parameters α and λ if it has density

$$f_X(t) = \begin{cases} \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Note that by definition, $\text{Gamma}(1, \lambda) = \text{Exponential}(\lambda)$.
If $X \sim \text{Gamma}(\alpha, \lambda)$,

$$\begin{aligned}
M_X(z) &= E(e^{zX}) \\
&= \int_0^\infty e^{zt} \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)} \, dt \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-(\lambda-z)t} \, dt \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{y}{\lambda - z} \right)^{\alpha-1} e^{-y} \frac{1}{\lambda - z} \, dy && y = (\lambda - z)t
\end{aligned}$$

Assume $\lambda - z > 0$, so $z < \lambda$.

$$\begin{aligned}
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{1}{(\lambda - z)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y} \, dy && y = (\lambda - z)t \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)(\lambda - z)^\alpha} \Gamma(\alpha) \\
&= \left(1 - \frac{z}{\lambda}\right)^{-\alpha} (z < \lambda)
\end{aligned}$$

Then

$$E(X) = M_X'(0) = \frac{\alpha}{\lambda}$$

$$\text{Var}(X) = M_X''(0) - (M_X'(0))^2 = \frac{\alpha}{\lambda^2}$$

Proposition. If $X_1, \ldots, X_k \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$, then $X_1 + \ldots + X_k \sim \text{Gamma}(k, \lambda)$.

Proof. By the theorem above,

$$M_{X_1 + \ldots X_k}(z) = \prod_{i=1}^k M_{X_i}(z) = (M_{X_1}(z))^k = \left(1 - \frac{z}{\lambda}\right)^{-k}$$

which is the mgf of a $\text{Gamma}(k, \lambda)$ random variable. ■

3 Conditioning

3.1 Conditional Expectation

Definition. The conditional expectation of a random variable X given an event A is

$$E(X \mid A) = \frac{E(XI(A))}{P(A)}$$

so long as $P(A) > 0$.

Theorem. $E(X \mid A)$ satisfies the axioms of expectation:

1. $E(1 \mid A) = 1$
2. $E(c_1X_1 + c_2X_2 \mid A) = c_1E(X_1 \mid A) + c_2E(X_2 \mid A)$
3. If $X \geq 0$, then $E(X \mid A) \geq 0$
4. If $X_n \uparrow X$, then $E(X_n \mid A) \uparrow E(X \mid A)$

Theorem (Law of Total Expectation)**.** If A_i are disjoint and $\bigcup_{i=1}^n A_i = \Omega$, then

$$E(X) = \sum_{i=1}^n P(A_i)E(X \mid A_i)$$

Proof. By definition of conditional expectation,

$$\begin{aligned}\sum_{i=1}^n P(A_i)E(X \mid A_i) &= \sum_{i=1}^n P(A_i)\frac{E(XI(A))}{P(A_i)} \\ &= E\left(X\sum_{i=1}^n I(A_i)\right) \\ &= E(X) && \text{since} \sum_{i=1}^n I(A_i) = 1\end{aligned}$$

■

Suppose X and Y are random variables and suppose Y is discrete. We can then define $E(X \mid Y = y)$ for all values y that Y takes. Generally, if Y takes on values $y_1, \ldots, y_n$, then we can calculate $E(X \mid Y = y_i) = \mu_i$. Define a random variable $Z = \mu_i$ with probability $P(Y = y_i)$. Then Z is the conditional expectation of X given Y.

- $Z = E(X \mid Y) = H(Y)$ where $H(y_i) = \mu_i$
- This means $E(X \mid Y)$ is a random variable and reflects the variability of X among different values of Y

In general, for all $A \subseteq \mathbb{R}$ such that $P(Y \in A) > 0$, $E(X \mid Y \in A)$ is well-defined. On the other hand, $E(X \mid Y)$ is a function $G(Y)$, so it must hold that

$$E(X \mid Y \in A) = E[G(Y) \mid A]$$

which implies

$$E[XI(Y \in A)] = E[G(Y)I(Y \in A)] \iff E[(X - G(Y))I(Y \in A)] = 0$$

by linearity. Since any function $H(Y)$ can be approximated by indicator functions, then

$$E((X - G(Y))H(Y)) = 0$$

This leads us to define conditional expectation over random variables as the following:

Definition. Let X and Y be random variables. The expectation of X conditional on Y, denoted $E(X \mid Y)$ is any solution $G(Y)$ satisfying

$$E[(X - G(Y))H(Y)] = 0 \tag{3}$$

for all functions H.

Theorem. The following hold:

(i) The definition is consistent with the definition in the discrete case.

(ii) If $E(X) < \infty$, then $E(X \mid Y)$ minimizes $D = E[(X - \varphi(Y))^2]$

(iii) Uniqueness: If $E(X^2) < \infty$, the solutions to (3) are almost surely equal

- i.e.: If $G_1(Y)$ and $G_2(Y)$ are solutions, then $G_1(Y) = G_2(Y)$ almost surely

Proof. (i) Recall that if Y is discrete and takes on values $y_1, \ldots, y_k$, then $G^D(y_i) = E(X \mid Y = y_i)$. We want to show $G^D(Y)$ is a solution to (3). If Y is discrete, then for any function H, we have

$$H(Y) = \sum_{i=1}^{n} I(Y = y_i)H(y_i)$$

thus it suffices to show for $H(Y) = I(Y = y_i)$ for all $i \in \{1, \ldots, k\}$. For all i, we want

$$E[XI(Y = y_i)] = E[G^D(Y)I(Y = y_i)]$$

Indeed, by definition of expectation conditional on the event $\{Y = y_i\}$,

$$\begin{aligned} E[G^D(Y)I(Y = y_i)] &= G^D(Y)E[I(Y = y_i)] \\ &= E[X \mid Y = y_i]P(Y = y_i) \\ &= E[XI(Y = y_i)] \end{aligned}$$

as required.
(ii) Let $G(Y) = E(X \mid Y)$ and define $\varphi^*(Y) = G(Y) - \varphi(Y)$. Then

$$\begin{aligned} D &= E[(X - G(Y) + G(Y) - \varphi(Y))^2] \\ &= E[(X - G(Y))^2] + 2E[(X - G(Y))\varphi^*(Y)] + E[(\varphi^*(Y))^2] \\ &= E[(X - G(Y))^2] + E[(\varphi^*(Y))^2] \qquad (4) \\ &\geq E[(X - G(Y))^2] \end{aligned}$$

as required.
(iii) Let $G_1(Y)$ and $G_2(Y)$ be solutions. Let $G_1(Y)$ be $\varphi(Y)$ in (4), so

$$E[(X - G_1(Y))^2] = E[(X - G_2(Y))^2] + E[(G_1(Y) - G_2(Y))^2] \tag{5}$$

Similarly,

$$E[(X - G_2(Y))^2] = E[(X - G_1(Y))^2] + E[(G_2(Y) - G_1(Y))^2] \tag{6}$$

Equations (5) and (6) imply that

$$E[(G_1(Y) - G_2(Y))^2] = 0$$

which implies $G_1(Y) - G_2(Y) = 0$ almost surely, as required. ■

Theorem. $E(X \mid Y)$ satisfies the axioms of expectation in an almost surely fashion:

1. If $X \geq 0$, then $E(X \mid Y) \geq 0$ almost surely
2. $E(1 \mid Y) = 1$ almost surely
3. $E(a_1 X_1 + a_2 X_2 \mid Y) = a_1 E(X_1 \mid Y) + a_2 E(X_2 \mid Y)$ almost surely for all a_1, a_2
4. If $X_i \uparrow X$, then $E(X_i \mid Y) \uparrow E(X \mid Y)$ almost surely

Properties of conditionals:

1. $E[L(Y)X \mid Y] = L(Y)E(X \mid Y)$ almost surely
2. Tower Law: $E[E(X \mid Y_1, Y_2) \mid Y_2] = E(X \mid Y_1)$
   - This implies $E[E(X \mid Y)] = E(X)$
3. If $E(X \mid Y_1, Y_2, Y_3)$ is a function $\psi(Y_1)$ of only Y_1, then

$$\psi(Y_1) = E(X \mid Y_1) = E(X \mid Y_1, Y_2)$$

4. Conditional decomposition of variance:
   Define $\text{Var}(X \mid Y) = E(X^2 \mid Y) - [E(X \mid Y)]^2$. Then

$$\text{Var}(X) = E[\text{Var}(X \mid Y)] + \text{Var}[E(X \mid Y)]$$

3.2 Conditional Probability

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

Proof.

$$P(A \mid B) = E[I(A) \mid B] = \frac{E[I(A)I(B)]}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

since $I(A)I(B) = I(A \cap B)$. ■

Properties:

1. $P(A \mid B) = \frac{P(A)}{P(B)} P(B \mid A)$

2. Law of Total Probability:

$$P(A) = \sum_{i=1}^{n} P(B_i) P(A \mid B_i)$$

where B_i are disjoint events with $\bigcup_{i=1}^{n} B_i = \Omega$

3. Let B_i's be defined as above. Then

$$P(B_i \mid A) = \frac{P(A \mid B_i) P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j) P(B_j)}$$

Proof. 1.

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

2. Follows from property of conditional expectation on the random variable $I(A)$.
3. Follows immediately from 1. and 2.. ■

3.3 Independence from a Conditional Perspective

Theorem. Two random variables X and Y are independent if, and only if, $E[H(X) \mid Y] = E[H(X)]$ almost surely for any H such that $E[H^2(X)] < \infty$.

Proof. To show sufficiency, by definition,

$$E[(X - E(X \mid Y))F(Y)] = 0$$

for all functions F so we only have to show

$$E[(H(X) - E[H(X)])F(Y)] = 0$$

holds. By independence,

$$E[(H(X) - E[H(X)])F(Y)] = E[H(X) - E[H(X)]]E[F(Y)] = 0$$

thus $E[H(X)] = E[H(X) \mid Y]$ almost surely.
To show necessity, for all functions $G(Y)$,

$$E[H(X) \mid Y]G(Y) = E[H(X)G(Y) \mid Y] \quad a.s.$$

thus we have

$$E[H(X)]G(Y) = E[H(X)G(Y) \mid Y] \quad a.s.$$

Taking expectation of both sides,

$$E[H(X)]E[G(Y)] = E[H(X)G(Y)]$$

which shows X and Y are independent by definition. ■

4 Continuous Random Variables and Their Transformations

4.1 Distributions with a Density

Definition. If $X = \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}^T$ is a random vector, then X is a continuous random vector if there exists a function $f(x_1, \ldots, x_n)$ such that $f \geq 0$ and

$$E[H(X)] = \int_{\mathbb{R}^n} H(x)f(x)\,dx$$

- Note that by axiom of expectation this implies that $\int_{\mathbb{R}^n} f(x)\,dx = 1$
- f is called the density function of X

Corollary. The following properties of density functions hold:

1. $P(X \in A) = \int_A f(x)\,dx$
2. Define the cdf of X as $F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$. Then

$$f(x_1, \ldots, x_n) = \frac{\partial F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}$$

3. For $r \leq n$, the density of $\begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}$ is given by

$$f(x_1, \ldots, x_r) = \int \cdots \int f(x_1, \ldots, x_n)\,dx_{r+1}dx_{r+2} \cdots dx_{n-1}dx_n$$

Theorem. If $X = \begin{bmatrix} X_1 & \cdots & X_n \end{bmatrix}$ is continuous with pdf $f(x_1, \ldots, x_n)$, then then X_i are independent iff

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n)$$

where $f_i(x_i)$ is the density of X_i.

4.1.1 Transformations

Suppose X and Y are random vectors with dimension m and r respectively with $r \leq qm$ and $Y = a(X)$. Suppose Y can be complemented by a transformation $Z = b(X)$ of dimension $m - r$ such that

$$X \to (Y, Z)$$

is an injective transformation and invertible with Jacobian

$$J(Y,Z) = \left|\det\left[\frac{\partial X}{\partial Y \partial Z}\right]\right|$$

Then the joint density of (Y,Z) is

$$f(x(y,z))J(y,z)$$

Consequently, the density of Y is

$$\int_{\mathbb{R}^{m-r}} f(x(y,z))J(y,z)\,dz$$

Proof. Note that

$$P(X \in A) = \int_A f(x)\,dx$$

Performing a change of variable by letting $(Y,Z) = (a(X), b(X))$, we have

$$P\left(\begin{bmatrix} Y \\ Z \end{bmatrix} \in c(A)\right) = \int_{c(A)} f(x(y,z))J(y,z)\,dydz$$

where $c(X) = (a(X), b(X)) = (Y,Z)$. This implies that the density of (Y,Z) is

$$f(x(y,z))J(y,z)$$

as required. ■

- Note that if we have $\begin{bmatrix} Y \\ Z \end{bmatrix} = AX$ where A is some invertible matrix, then $J(y,z) = \left|\frac{1}{\det(A)}\right|$

4.2 Conditional Densities

Theorem. Suppose X and Y are continuous random vectors with joint density $f(x,y)$. The distribution of X conditional on Y has density

$$f(x \mid y) = \frac{f(x,y)}{f_Y(y)} \tag{7}$$

where $f_Y(y) = \int f(x,y)\,dy$ is the density of Y.

Proposition. The definition of the conditional density is consistent with the definition of conditional expectation.

Proof. By (7), we have

$$E[H(X) \mid Y] = \int H(x)f(x \mid y)\,dx$$

By definition, $E[H(X) \mid Y]$ should satisfy

$$E[H(X)G(Y)] = E[E[H(X) \mid Y]G(Y)]$$

for all functions G. Since $E[H(X) \mid Y]$ is a function of Y, then

$$\begin{aligned}
E[E[H(X) \mid Y]G(Y)] &= \int E[H(X) \mid Y = y]G(Y)f_Y(y)\,dy \\
&= \int \left[\int H(x)f(x \mid y\,dx\right] G(y)f_Y(y)\,dy \\
&= \iint H(x)f(x \mid y)f_Y(y)G(y)\,dxdy \\
&= \iint H(x)f(x,y)G(y)\,dxdy
\end{aligned}$$

On the other hand,

$$E[H(X)G(Y)] = \iint H(x)G(y)f(x,y)\,dxdy$$

which shows equality, as desired. ■

4.3 Order Statistics

Suppose $X_1, \ldots, X_n$ are iid random variables with pdf $f(x)$ and cdf $F(x)$. We can order the X_i

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

where $X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

4.3.1 Distribution of Order Statistics

For all $x \in \mathbb{R}$,

$$\begin{aligned}
P(X_{(n)} \leq x) &= P(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) \\
&= \prod_{i=1}^{n} P(X_i \leq x) && \text{by independence} = [F(x)]^n
\end{aligned}$$

so the density function of $X_{(n)}$ is $n[F(x)]^{n-1}f(x)$. On the other hand,

$$\begin{aligned}
P(X_{(1)} \leq x) &= 1 - P(X_1 > x, \ldots, X_n > x) \\
&= 1 - [1 - F(x)]^n
\end{aligned}$$

thus the density function of $X_{(1)}$ is $n[1 - F(x)]^{n-1}f(x)$.
For any $X_{(i)}$, consider

$$\frac{P(x \leq X_{(i)} \leq x + dx}{dx}$$

where dx is very small. By definition of the order statistics, $P(x \leq X_{(i)} \leq x + dx)$ is the same as the probability that $i-1$ X_j's must be $\leq x$, one of them is between x and dx, and the rest are greater than $x + dx$, which shows

$$\begin{aligned} P(x \leq X_{(i)} \leq x + dx) &= \binom{n}{i-1}[F(x)]^{i-1}\binom{n-i+1}{1}[F(x+dx) - F(x)][1 - F(x) + dx]^{n-i} \\ &= \binom{n}{i-1}(n-i+1)[F(x)]^{i-1}f(x)dx[1 - F(x+dx)]^{n-i} \end{aligned}$$

since dx is small. This means

$$\frac{P(x \leq X_{(i)} \leq x + dx)}{dx} = \binom{n}{i-1}(n-i+1)[F(x)]^{i-1}f(x)[1 - F(x+dx)]^{n-i}$$

Let $dx \to 0$. Then

$$f_{X_{(i)}}(x) = \binom{n}{n-1}(n-i+1)[F(x)]^{i-1}f(x)[1 - F(x)]^{n-i}$$

5 Basic Limit Theorems

5.1 Convergence in Probability

Definition. Let $X_1, \ldots, X_n$ be a sequence of random variables. $X_i \to X$ in probability if

$$\lim_{i\to\infty} P(|X_i - X| > \varepsilon) = 0$$

for all $\varepsilon = 0$.

Proposition. Suppose X_n is a sequence of random variables.

1. If $X_n \xrightarrow{p} X$, then $cX_n \xrightarrow{p} cX$ for constant c.
2. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_n + Y_n \xrightarrow{p} X + Y$
3. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $X_nY_n \xrightarrow{p} XY$
4. If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} c$ where $c \neq 0$ is a constant, then $\frac{X_n}{Y_n} \xrightarrow{p} \frac{X}{c}$

Proof. Of 3. For all $\varepsilon > 0$, we have $P(|X_n - X| > \varepsilon) \to 0$ and $P(|Y_n - Y| > \varepsilon) \to 0$. Note that

$$X_nY_n - XY = (X_n - X)Y)n + X(Y_n - Y) = (X_n - X)(Y_n - Y) + Y(X_n - X) + X(Y_n - Y)$$

Thus for all $\varepsilon > 0$,

$$P(|X_nY_n - XY| > \varepsilon) \leq P\left(|(X_n - X)(Y_n - Y)| > \frac{\varepsilon}{3}\right) + P\left(|Y(X_n - X)| > \frac{\varepsilon}{3}\right) + P\left(|X(Y_n - Y)| > \frac{\varepsilon}{3}\right)$$

Denote the 3 expressions on the RHS as 1, 2, 3 respectively. We claim $1, 2, 3 \to 0$ as $n \to \infty$.
For 1, WLOG let $\varepsilon < 1$. Note that if $P(|X_n - X|> \varepsilon)$, then $P(|X_n - X|> \delta) \to 0$ for all $\delta > \varepsilon$ as $P(|X_n - X|> \delta) \leq P(|X_n - X|> \varepsilon)$ if $\delta > \varepsilon$. By assumption, $P(|X_n - X|> 1) \to 0$ as $n \to \infty$ and $P\left(|Y_n - Y|> \frac{\varepsilon}{3}\right) \to 0$). For any $\delta > 0$, there exists some $N_\delta \in \mathbb{N}$ such that for all $n \geq N_\delta$,

$$P(|X_n - X|> 1) \leq \frac{\delta}{8}$$
$$P\left(|Y_n - Y|> \frac{\varepsilon}{3}\right) \leq \frac{\delta}{8}$$

which implies $P\left(|X_n - X||Y_n - Y|> \frac{\varepsilon}{3}\right) \leq \frac{\delta}{4}$ if $n \geq N_\delta$.
For 2 and 3, we claim that $P(|X|\geq M) \to 0$ if $M \to \infty$. Since for all $\omega \in \Omega$, $|X(\omega)|< M$ if M is large enough given that $X_n \to X$ where $|X|< \infty$, then $I(|X|\geq M) \to 0$ as $M \to \infty$. Then since $0 \leq I(|X|\geq M) \leq 1$, by DCT, $E(I(|X|\geq M)) \to E(0) = 0$ as $M \to \infty$. So, there exists some $M_\delta \in \mathbb{N}$ such that $P(|X|\geq M_\delta) \leq \frac{\delta}{8}$. By assumption, there exists some N^*_δ such that $P\left(|Y_n - Y|> \frac{\varepsilon}{3}\frac{1}{M_\delta}\right) \leq \frac{\delta}{8}$ for $n > N^*_\delta$. Thus,

$$P\left(|X||Y_n - Y|> \frac{\varepsilon}{3}\frac{1}{M_\delta}M_\delta\right) \leq \frac{\delta}{4}$$

if $n > N^*$. We apply a similar argument to $|Y||X_n - X|$. So, if n is sufficiently large,

$$P(|X_nY_n - XY|> \varepsilon) \leq \frac{\delta}{4} + \frac{\delta}{4} + \frac{\delta}{4} < \delta$$

which shows $P(|X_nY_n - XY|> \varepsilon) \to 0$, as required. ■

Proposition. For some $r > 0$, if $E(|X_n - X|^r) \to 0$ as $n \to \infty$, then $X_n \xrightarrow{p} X$.

Proof. We need $P(|X_n - X|> \varepsilon) \to 0$ for all $\varepsilon > 0$. Note that

$$P(|X_n - X|> \varepsilon) = P(|X_n - X|^r > \varepsilon^r) \leq \frac{E(|X_n - X|^r)}{\varepsilon^r} \to 0$$

by Markov's Inequality. ■

Corollary. If $E[(X_n - X)^2] \to 0$, then $X_n \xrightarrow{p} X$.

5.1.1 Weak Law of Large Numbers

If $X_1, X_2, \ldots$ is a sequence of random variables with $E(X_i) = \mu_i$, $\mathrm{Var}(X_i) = \sigma^2 > 0$ and $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

Proof. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then

$$E[(\bar{X}_n - \mu)^2] = E\left[\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)\right)^2\right]$$

Let $Y_i = X_i - \mu$, so $E(Y_i) = 0$, $E(Y_i^2) = \sigma^2$, and $\text{Cov}(Y_i, Y_j) = 0$ for all $i \neq j$. Then

$$\begin{aligned} E[(\bar{X}_n - \mu)^2] &= \frac{1}{n^2} E\left[\left(\sum_{i=1}^n Y_i\right)^2\right] \\ &= \frac{1}{n^2}\left(\sum_{i=1}^n E(Y_i^2) + \sum_{i \neq j} E(Y_i Y_j)\right) \\ &= \frac{\sigma^2}{n} \to 0 \end{aligned}$$

as required. ■

Theorem. If $X_1, \ldots, X_n$ are independent with cdf F and the empirical cdf is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$$

then $F_n(x) \xrightarrow{p} F(x)$ for all x.

Proof. Note that for all x, we have

$$E[I(X_i \leq x)] = P(X_i \leq x) = F(x)$$

Since each X_i is independent, then so is each $I(X_i \leq x)$, thus $\text{Cov}[I(X_i \leq x), I(X_j \leq x)] = 0$ for all $i \neq j$. By the WLLN, we have

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x) \xrightarrow{p} F(x)$$

as required. ■

5.2 Convergence in Distribution

Definition. A sequence of random variables $X_1, X_2, \ldots$ converges in distribution to X if

$$E[H(X_n)] \to E[H(X)]$$

for any bounded and continuous H.

- Notice how this definition does not require X_n to be close to X
- $\xrightarrow{d}$ does not imply $\xrightarrow{p}$

Theorem. $X_n \xrightarrow{d} X$ iff $P(X_n \leq x) \to P(X_n \leq x)$ at any point x at which the cdf of X is continuous.

Theorem. $X_n \xrightarrow{d} X$ if $M_{X_n}(t) \to M_X(t)$ as $n \to \infty$ for all t in a neighbourhood of 0.

5.2.1 Normal Random Variables

Definition. X is a normal random variable with mean μ and variance σ^2 if it has density

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

- The MGF of $\mathcal{N}(0,1)$ is $\exp(\frac{t^2}{2})$

5.2.2 Central Limit Theorem

Let $X_1, \ldots$ be iid with mean μ and variance $\sigma^2 < \infty$. Let $Y_n = \frac{\sqrt{n}(X_n - \mu)}{\sigma}$ where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $Y_n \xrightarrow{d} \mathcal{N}(0,1)$.

Theorem. $\xrightarrow{p}$ implies $\xrightarrow{d}$.