Math 170A Running Notes

Brendan Connelly

January to March 2025

1 Basics and Miscellaneous

Definition (Basic Set Theory)**.** Let $h : \Omega_1 \to \Omega_2$ and $A \subseteq \Omega_2$. The preimage of A under h is defined as:

$$h^{-1}(A) = \{\omega \in \Omega_1 \mid h(\omega) \in A\}.$$

For an arbitrary collection $\{A_i\}_{i \in I}$ of subsets $A_i \subseteq \Omega_2$, we have:

$$h^{-1}\left(\bigcup_{i \in I} A_i\right) = \bigcup_{i \in I} h^{-1}(A_i) \quad \text{and} \quad h^{-1}\left(\bigcap_{i \in I} A_i\right) = \bigcap_{i \in I} h^{-1}(A_i).$$

For the image of a set $B \subseteq \Omega_1$, defined as:

$$h(B) := \{\omega' \in \Omega_2 \mid \omega' = h(\omega) \text{ for some } \omega \in B\},$$

we have:

$$h\left(\bigcup_{i \in I} B_i\right) = \bigcup_{i \in I} h(B_i) \quad \text{and} \quad h\left(\bigcap_{i \in I} B_i\right) \subseteq \bigcap_{i \in I} h(B_i),$$

with equality in the second case if h is injective.

Similarly, DeMorgan's laws hold for arbitrary collections of $A_i \subseteq \Omega$:

$$\left(\bigcup_{i \in I} A_i\right)^c = \bigcap_{i \in I} A_i^c \quad \text{and} \quad \left(\bigcap_{i \in I} A_i\right)^c = \bigcup_{i \in I} A_i^c,$$

where we define $A^c = \Omega \setminus A = \{\omega \in \Omega \mid \omega \notin A\}$.

Definition (Axiom of Choice)**.** The *Axiom of Choice* states that given any collection $\{X_i\}_{i \in I}$ of nonempty sets indexed by a set I, there exists a function $f : I \to \bigcup_{i \in I} X_i$ such that $f(i) \in X_i$ for all $i \in I$. This function f is called a *choice function*.

Theorem (Banach-Tarski Paradox)**.** Let S be a solid ball in $\mathbb{R}^3$. It is possible to partition S into a finite number of disjoint subsets, $S_1, S_2, \ldots, S_n$, such that these subsets can be reassembled (using only rotations and translations) into two solid balls, each congruent to the original ball S.

This paradox relies on the Axiom of Choice and demonstrates that certain notions of volume and measure fail in higher-dimensional spaces.

Definition (Borel Set)**.** A *Borel set* is any set in a σ-algebra generated by the open subsets of a topological space X.

More formally, let $\mathcal{T}$ denote the collection of open sets in X. The *Borel σ-algebra*, denoted $\mathcal{B}(X)$, is the smallest σ-algebra containing all the open sets in X. A set $A \subseteq X$ is a Borel set if $A \in \mathcal{B}(X)$.

2 σ-Algebras

Definition (Sample Space)**.** The *sample space* (Ω) is the set of all possible outcomes of a probabilistic experiment.

For example, for the experiment of flipping two coins:

$$\Omega = \{\text{TT, TH, HT, HH}\}.$$

Definition (Outcome)**.** An *outcome* (ω) is an element of the sample space Ω.

For example, $\omega \in \Omega$ could be HT in the experiment of flipping two coins.

Definition (Event)**.** An *event* is a subset of the sample space, $A \subseteq \Omega$.

For example, the event of getting at least one head in a coin flip is:

$$A = \{\text{HT, TH, HH}\}.$$

Definition (σ-Algebra)**.** A σ-algebra on a sample space Ω is a non-empty collection $\mathcal{F}$ of subsets $A \subseteq \Omega$ satisfying the following properties:

1. **Complement Closed:** If $A \in \mathcal{F}$, then $A^c = \Omega \setminus A \in \mathcal{F}$.
2. **Countable Union Closed:** If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A set Ω equipped with a σ-algebra $\mathcal{F}$ is called a ***measurable space***. These are the subsets that can be assigned a measure or probability.

Proposition. Let $\mathcal{F}$ be a σ-algebra on Ω. Then:

1. $\Omega \in \mathcal{F}$ and $\emptyset \in \mathcal{F}$.
2. **(Countable Intersection Closed)** If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$.

Remark. The intersections and unions in a σ-algebra can be taken to be finite.

Example (Examples of σ-Algebras)**.**

- The power set $\mathcal{P}(\Omega)$ is a σ-algebra, often called the "finest" σ-algebra.
- The collection $\{\emptyset, \Omega\}$ is a σ-algebra, often called the "coarsest" σ-algebra.
- The collection $\{\emptyset, A, A^c, \Omega\}$ for a single $A \subseteq \Omega$ is a σ-algebra.
- A topology on Ω determines a coarsest σ-algebra containing it.

Definition (Measurable Function)**.** Let $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ be measurable spaces. A function $h : \Omega_1 \to \Omega_2$ is *measurable* if:

$$h^{-1}(A) \in \mathcal{F}_1 \quad \text{for all } A \in \mathcal{F}_2.$$

Definition (Generated σ-Algebra)**.** Let $(E, \mathcal{E})$ be a measurable space and $X : \Omega \to E$. We define the σ-algebra on Ω generated by X as:

$$\mathcal{F}_X := \{X^{-1}(A) \mid A \in \mathcal{E}\}.$$

This is the coarsest σ-algebra with respect to which X is measurable.

Thus, X is measurable with respect to a σ-algebra $\mathcal{F}$ on Ω if and only if $\mathcal{F}_X \subseteq \mathcal{F}$.

Definition (Measure)**.** Let $(E, \mathcal{E})$ be a measurable space. A *measure* μ on $(E, \mathcal{E})$ is a function $\mu : \mathcal{E} \to [0, \infty]$ satisfying the following:

1. (**Non-negativity**) For each $A \in \mathcal{E}$, $\mu(A) \geq 0$ (including ∞).
2. (**Countable Additivity**) For $A_1, A_2, \dots \in \mathcal{E}$ such that $A_i \cap A_j = \emptyset$ for $i \neq j$,
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$
3. $\mu(\emptyset) = 0$ (implied if there exists $A \in \mathcal{E}$ of finite measure).

Remark. The union/sum in the above can be taken to be finite.

Definition (Measure Space)**.** A *measure space* $(E, \mathcal{E}, \mu)$ is a measurable space $(E, \mathcal{E})$ equipped with a measure μ. It is called *discrete* if E is countable and $\mathcal{E} = \mathcal{P}(E)$, the power set of E.

3 Probability Spaces

Definition (Probability Space)**.** A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ is a measure space where $\mathbb{P}(\Omega) = 1$.

For $A \in \mathcal{F}$, $\mathbb{P}(A)$ is interpreted as the probability that A occurs, or equivalently, that $\omega \in A$.

Proposition (Basic Properties of Probability Spaces)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then the following hold:

1. For $A \in \mathcal{F}$,
$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1.$$
2. For $A, B \in \mathcal{F}$,
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$
3. For $A \subseteq B$, $A, B \in \mathcal{F}$,
$$\mathbb{P}(A) \leq \mathbb{P}(B).$$
4. Let Ω be a countable set. Show that $\mathcal{F} = \mathcal{P}(\Omega)$ is the unique σ-algebra containing the singleton sets $\{\omega\} \subseteq \Omega$ for all $\omega \in \Omega$.
5. Let Ω be countable and $\mathcal{F} = \mathcal{P}(\Omega)$. Show that $\mathbb{P}$ is uniquely determined by the convergent sequence $\{\mathbb{P}(\{\omega\})\}_{\omega \in \Omega}$ as:
$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) \quad \text{for any } A \subseteq \Omega.$$

Remark. Properties 4 and 5 both hold for Ω finite as well.

Definition (Probability: New Spaces from Old - Uniform Probability Measure)**.** Let Ω be a finite set. The *uniform probability measure* is defined by

$$\mathbb{P}(\{\omega\}) = \frac{1}{N},$$

for every $\omega \in \Omega$, where $N = |\Omega|$ is the cardinality of Ω.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(E, \mathcal{E})$ a measurable space, and $X : \Omega \to E$ measurable, called an *E-valued random variable*. Define the *pushforward probability measure* $\mathbb{P}_X$ on $(E, \mathcal{E})$ by:

$$\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}),$$

for all $A \in \mathcal{E}$.

Definition (Probability: New Spaces from Old)**.** Let Ω be a sample space and $A, B \subseteq \Omega$ be events with $\mathbb{P}(B) > 0$. The *conditional probability* $\mathbb{P}(A \mid B)$ is defined by:

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

Proposition. $\mathbb{P}(\cdot \mid B) : \mathcal{F} \to \mathbb{R}$ is a probability measure on $(\Omega, \mathcal{F})$.

Remark. We can restrict the σ-algebra $\mathcal{F}$ to define a σ-algebra on B by $\mathcal{F}_B = \{A \cap B \mid A \in \mathcal{F}\}$. We can equivalently view $\mathbb{P}(\cdot \mid B)$ as a measure on $(B, \mathcal{F}_B)$.

Remark (Multiplication Rule)**.** For events A, B, C,

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

and more generally:

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B \mid A)\mathbb{P}(C \mid A \cap B).$$

Independence

Definition. Two events $A, B \subseteq \Omega$ are *independent* if:

$$\mathbb{P}(A \mid B) = \mathbb{P}(A),$$

or equivalently:

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

Definition. Two E_i-valued random variables $X_i : \Omega \to (E_i, \mathcal{E}_i)$ are *independent* if $X_1^{-1}(A)$ and $X_2^{-1}(B)$ are independent for every $A \in \mathcal{E}_1$ and $B \in \mathcal{E}_2$.

Example. Two random variables $X_1, X_2 : \Omega \to \mathbb{Z}$ are independent if for all $x_1, x_2 \in \mathbb{Z}$,

$$\mathbb{P}(X_1 = x_1 \text{ and } X_2 = x_2) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2).$$

Definition (Mutual Independence of More Than Two Events)**.** Given a collection of more than two events $\{A_i\}_{i \in I}$ or random variables, we say they are *mutually independent* if:

1. They are pairwise independent, meaning for all $i, j \in I$ with $i \neq j$,

$$P(A_i \cap A_j) = P(A_i)P(A_j).$$

2. The probabilities of all possible intersections of subsets of the events factor as the products of their individual probabilities. That is, for any subset $\{i_1, i_2, \ldots, i_k\} \subseteq I$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

For example, for three events A, B, and C to be mutually independent, the following conditions must hold:

$$P(A \cap B) = P(A)P(B), \quad P(B \cap C) = P(B)P(C), \quad P(A \cap C) = P(A)P(C),$$

and

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

Mutual independence extends pairwise independence by requiring the same factorization property for intersections of more than two events.

3.1 Bayes

Theorem (Basic Bayes' Rule)**.** Let A and B be events with $P(A), P(B) > 0$. The relationship between their probabilities is given by:

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

In its simplest form, Bayes' rule follows directly:

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

This formula allows us to calculate $P(B \mid A)$ (the posterior probability) using $P(A \mid B)$ (the likelihood), $P(A)$ (the evidence), and $P(B)$ (the prior probability).

Bayes' rule is a method for *updating beliefs*:

- $P(B)$ is the initial assumption or *prior* probability of B.
- $P(B \mid A)$ is the *posterior* probability of B after observing A.
- The factor $\frac{P(A|B)}{P(A)}$ is the *re-weighting factor*, which adjusts the prior probability based on the new evidence A.

The re-weighting factor $\frac{P(A|B)}{P(A)} > 1$ if and only if $P(A \mid B) > P(A)$, which occurs when B makes A more likely.

Theorem (Law of Total Probability)**.** Let $B_1, \dots, B_n \subseteq \Omega$ be a partition of the sample space Ω. For any event $A \subseteq \Omega$, we have:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n),$$

or equivalently,

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n).$$

pf:

$$\begin{aligned}
P(A) &= P\left(A \cap \bigcup_i B_i\right) \\
&= P\left(\bigcup_i (A \cap B_i)\right) \\
&= \sum_i P(A \cap B_i) && \text{(by 1.14)} \\
&= \sum_i P(A \mid B_i)P(B_i) && \text{(by 1.32).}
\end{aligned}$$

Theorem (Extended Bayes' Rule)**.** Let $B_1, \dots, B_n \subseteq \Omega$ be a partition of the sample space Ω. For any event $A \subseteq \Omega$ and any $i \in \{1, \dots, n\}$, we have:

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

or equivalently,

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \cdots + P(A \mid B_n)P(B_n)}.$$

This result expresses $P(B_i \mid A)$ as the portion of $P(A)$ contributed by $P(A \cap B_i)$.

4 Discrete Random Variables

Proposition (Measurability of a Function)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space. A function $X : (\Omega, \mathcal{F}) \to (E, \mathcal{E})$ is measurable if and only if the function $f : (\Omega, \mathcal{F}) \to (\mathrm{V}_X, \mathcal{E}|_{\mathrm{V}_X})$ is measurable, where $\mathrm{V}_X = \mathrm{im}(X) \subset E$ and

$$\mathcal{E}|_{\mathrm{V}_X} = \{A \cap \mathrm{V}_X \mid A \in \mathcal{E}\}.$$

Remark. For any countable $\mathrm{V}_X \subset \mathbb{R}$ and any σ-algebra $\mathcal{E}_{\mathbb{R}}$ on $\mathbb{R}$ containing $\{x\}$ for each $x \in \mathbb{R}$, we have

$$\mathcal{E}_{\mathbb{R}}|_{\mathrm{V}_X} = \mathcal{P}(\mathrm{V}_X).$$

Proposition (Measurability Criterion)**.** Let $X : \Omega \to \mathbb{R}$ be a function which takes at most countably many values $\mathrm{V}_X = \mathrm{im}(X) \subset \mathbb{R}$, and let $\mathcal{E}_{\mathbb{R}}$ be any σ-algebra on $\mathbb{R}$ as above. The function $X : \Omega \to \mathbb{R}$ is measurable if and only if

$$\{X = x\} = X^{-1}(\{x\}) \in \mathcal{F} \quad \text{for each } x \in \mathrm{V}_X = \mathrm{im}(X) \subset \mathbb{R}.$$

Definition (Discrete Random Variable)**.** A *discrete random variable* is a measurable function $X : \Omega \to \mathbb{R}$ which takes at most countably many values $\mathrm{V}_X \subset \mathbb{R}$. Measurability ensures that $\mathbb{P}(X = x)$ is defined for each $x \in \mathrm{V}_X$.

Remark. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a discrete random variable with a (countable) set of possible values $\mathrm{V}_X \subset \mathbb{R}$.

We can equivalently view the function $X : \Omega \to \mathrm{V}_X$, and thus V_X becomes a probability space with respect to

$$\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A)).$$

Definition (Probability Mass Function)**.** The *probability mass function* $f_X : \mathrm{V}_X \to [0, 1]$ is the function defined as

$$f_X(x) = \mathbb{P}_X(\{x\}) = \mathbb{P}(X^{-1}(\{x\})) = \mathbb{P}(\{X = x\}).$$

Since V_X is a discrete measurable space, specifying the measure $\mathbb{P}_X$ is equivalent to specifying the summable series $\{\mathbb{P}_X(\{x\})\}_{x \in \mathrm{V}_X}$ or equivalently the probability mass function $f_X : \mathrm{V}_X \to [0, 1]$.

Example (Binomial Distribution)**.** Let $\Omega = \{H, T\}^n$ be the sample space for n independent flips of a coin, where p is the probability of heads and $q = 1 - p$ is the probability of tails. Let $X : \Omega \to \mathbb{Z}$ represent the number of heads. Then

$$\mathrm{V}_X = \{0, 1, \ldots, n\},$$

and the probability mass function is given by

$$f_X(x) = \binom{n}{x} p^x q^{n-x}, \quad x \in \mathrm{V}_X\,.$$

Proposition (Geometric Distribution)**.** Let X be the number of independent coin flips required to obtain the first head, where the probability of heads is $p \in (0, 1)$ (and $q = 1 - p$ is the probability of tails). Then X takes values in $\{1, 2, 3, \ldots\}$ with probability mass function

$$P(X = k) = q^{k-1} p, \quad k = 1, 2, 3, \ldots.$$

In particular, note that

$$\sum_{k=1}^{\infty} P(X = k) = p \sum_{k=0}^{\infty} q^k = \frac{p}{1 - q} = \frac{p}{p} = 1,$$

so this indeed defines a valid probability distribution.

Proposition (Negative Binomial Distribution)**.** Let $r \in \mathbb{N}$ and consider the experiment of flipping a coin with probability $p \in (0, 1)$ of heads (and $q = 1 - p$ of tails) until the r^{th} head occurs. Define X to be the number of flips required. Then X takes values in $\{r, r+1, r+2, \dots\}$ and its probability mass function is given by

$$P(X = x) = \binom{x-1}{r-1} q^{x-r} p^r, \quad x = r, r+1, r+2, \dots.$$

In the special case $r = 2$, this formula becomes

$$P(X = x) = (x-1) q^{x-2} p^2, \quad x = 2, 3, 4, \dots.$$

Definition (Expected Value)**.** Let $X : \Omega \to \mathbb{R}$ be a random variable with range

$$V_X = \operatorname{im}(X) \subset \mathbb{R},$$

the set of possible values taken by X. The *expected value* of X is defined as

$$E(X) = \sum_{x \in V_X} x\, P(X = x),$$

whenever the above sum converges absolutely.

Proposition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a discrete random variable with range

$$V_X = \operatorname{im}(X) \subset \mathbb{R},$$

and with induced probability measure $\mathbb{P}_X$ on V_X. Consider any measurable function $h : \mathbb{R} \to \mathbb{R}$ (for example, a continuous function). Then the composition

$$Y = h \circ X : \Omega \to \mathbb{R}$$

defines a discrete random variable on Ω. In particular, the range of Y is

$$V_Y = h(V_X),$$

and Y induces a probability measure $\mathbb{P}_Y$ on V_Y by

$$\mathbb{P}_Y(A) = \mathbb{P}\big(Y^{-1}(A)\big), \quad \text{for all } A \subseteq V_Y.$$

Theorem. Using the notation above, consider the restriction

$$h|_{V_X} : V_X \to \mathbb{R},$$

which may be viewed as a random variable on the probability space $(V_X, \mathcal{P}(V_X), \mathbb{P}_X)$. This random variable induces a measure $\tilde{\mathbb{P}}_Y$ on V_Y. Then the measures $\mathbb{P}_Y$ and $\tilde{\mathbb{P}}_Y$ agree. In particular, the expected value of $h(X)$ is given by

$$E\big(h(X)\big) = \sum_{x \in V_X} h(x)\, \mathbb{P}\big(X = x\big).$$

Proposition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a discrete random variable.

1. If $c(\omega) = c \in \mathbb{R}$ is a constant function, then

$$E(c) = c.$$

2. For any $c_1, c_2 \in \mathbb{R}$ and measurable functions $h_1, h_2 : \mathbb{R} \to \mathbb{R}$,

$$E\big(c_1 h_1(X) + c_2 h_2(X)\big) = c_1 E\big(h_1(X)\big) + c_2 E\big(h_2(X)\big).$$

Definition (Conditional Expectation)**.** Let $B \subseteq \Omega$ be an event with $\mathbb{P}(B) > 0$, so that the conditional probability $\mathbb{P}(\cdot \mid B)$ is well-defined. The expectation of X conditional on B is given by

$$E(X \mid B) = \sum_{x \in V_X} x \, \mathbb{P}(X = x \mid B).$$

Proposition. Let $\{B_i\}_{i=1}^n$ be a finite partition of Ω with $\mathbb{P}(B_i) > 0$ for each i. Then the expectation of X satisfies the law of total expectation:

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

Definition (Variance)**.** Variance measures how far the values of a random variable typically deviate from the mean $\mu = E(X)$, or geometrically, how "spread out" the probability mass function f_X is.

A naive guess for this measure might be $E(X - \mu)$, but since

$$E(X - \mu) = E(X) - E(\mu) = 0,$$

the positive and negative deviations cancel out. Instead, we define the *variance* of X as

$$\sigma_X^2 = E\big((X - \mu)^2\big).$$

This expression ensures that all deviations from μ are weighted positively, providing a meaningful measure of dispersion.

Proposition (Computing Variance)**.** For a discrete random variable X with probability mass function f_X,

$$E(h(X)) = \sum_{x \in V_X} h(x) \, \mathbb{P}(X = x).$$

Applying this to $h(X) = (X - \mu)^2$, we obtain

$$\sigma_X^2 = E\big((X - \mu)^2\big) = \sum_{x \in V_X} (x - \mu)^2 \mathbb{P}(X = x).$$

Definition (Poisson Distribution)**.** The *Poisson distribution* models the number of times an event occurs in a fixed time interval, assuming events occur independently and at a constant mean rate $\lambda > 0$. It arises as the limit of a binomial process where the number of sub-intervals tends to infinity while the probability of an event in each sub-interval tends to zero.

A random variable X follows a Poisson distribution with parameter λ if

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots.$$

Proposition. The variance of a random variable X satisfies the identity:

$$\sigma^2(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2.$$

Proposition. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(E_0, \mathcal{E}_0)$, $(E_1, \mathcal{E}_1)$ measurable spaces. Suppose that

$$X : \Omega \to E_0 \quad \text{and} \quad h : E_0 \to E_1$$

are measurable. Then the composition

$$Y = h \circ X : \Omega \to E_1$$

is measurable and the induced measure $\mathbb{P}_Y$ on E_1 satisfies

$$\mathbb{P}_Y(A) = \mathbb{P}\Big(\{\omega \in \Omega : Y(\omega) \in A\}\Big) = \mathbb{P}\Big(\{\omega \in \Omega : X(\omega) \in h^{-1}(A)\}\Big)$$

for all $A \in \mathcal{E}_1$.

In particular, if $E_1 = \mathbb{R}$ and E_0 (and thus Y) is discrete, then

$$E(Y) = \sum_{y \in V_Y} y\, \mathbb{P}_Y(\{y\}) = \sum_{e \in E_0} h(e)\, \mathbb{P}_X(\{e\}).$$

Moreover, given discrete random variables $X_1, X_2 : \Omega \to \mathbb{R}$, define

$$X = (X_1, X_2) : \Omega \to V_{X_1} \times V_{X_2},$$

with

$$\mathbb{P}_X\big((x_1, x_2)\big) = \mathbb{P}\big(\{X_1 = x_1\} \cap \{X_2 = x_2\}\big).$$

Then for any measurable function $h : \mathbb{R}^2 \to \mathbb{R}$ we have

$$E\big(h(X_1, X_2)\big) = \sum_{(x_1, x_2) \in V_{X_1} \times V_{X_2}} h(x_1, x_2)\, \mathbb{P}\big(\{X_1 = x_1\} \cap \{X_2 = x_2\}\big).$$

Proposition (Basic Properties of Joint Distributions)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y : \Omega \to \mathbb{R}$ be discrete random variables. Define the joint probability mass function

$$f_{X,Y}(x, y) = \mathbb{P}\big(\{X = x\} \cap \{Y = y\}\big), \quad (x, y) \in V_X \times V_Y.$$

Then:

(i) X and Y are independent if and only if

$$f_{X,Y}(x, y) = f_X(x) f_Y(y)$$

for all $(x, y) \in V_X \times V_Y$.

(ii) For any function $h : \mathbb{R}^2 \to \mathbb{R}$, the expectation is given by

$$E\big(h(X, Y)\big) = \sum_{(x,y) \in V_X \times V_Y} h(x, y)\, f_{X,Y}(x, y) = \sum_{x \in V_X} \sum_{y \in V_Y} h(x, y)\, f_{X,Y}(x, y).$$

(iii) The marginal distribution of X is determined by

$$f_X(x) = \mathbb{P}\big(X = x\big) = \sum_{y \in V_Y} f_{X,Y}(x, y).$$

(iv) In particular, if $h(x,y) = g(x)$ depends only on x, then

$$E\big(g(X)\big) = \sum_{(x,y)\in V_X\times V_Y} g(x)\, f_{X,Y}(x,y) = \sum_{x\in V_X} g(x) \left(\sum_{y\in V_Y} f_{X,Y}(x,y) \right) = \sum_{x\in V_X} g(x)\, f_X(x).$$

Proposition (Properties of Expectation for Discrete Random Variables)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X, Y : \Omega \to \mathbb{R}$ be discrete random variables. Then:

1. For any measurable functions $g, h : \mathbb{R} \to \mathbb{R}$,
$$E\big(g(X) + h(Y)\big) = E\big(g(X)\big) + E\big(h(Y)\big).$$
2. If X and Y are independent, then
$$E(XY) = E(X)\,E(Y).$$
3. X and Y are independent if and only if for any measurable functions $g, h : \mathbb{R} \to \mathbb{R}$,
$$E\big(g(X)h(Y)\big) = E\big(g(X)\big)\,E\big(h(Y)\big).$$
4. If X and Y are independent, then
$$E\big((X-\mu_X)(Y-\mu_Y)\big) = 0,$$
and consequently,
$$\sigma^2(X+Y) = \sigma^2(X) + \sigma^2(Y).$$

Definition (Covariance)**.** Let X and Y be random variables with means $\mu_X = E(X)$ and $\mu_Y = E(Y)$. The *covariance* of X and Y is defined by

$$\sigma(X,Y) = E\big((X-\mu_X)(Y-\mu_Y)\big) = \sum_{(x,y)\in V_X\times V_Y} (x-\mu_X)(y-\mu_Y)\, f_{X,Y}(x,y),$$

where $f_{X,Y}(x,y) = \mathbb{P}(\{X=x\}\cap\{Y=y\})$. This measure is positive when X and Y tend to deviate above or below their means simultaneously, and negative when they deviate in opposite directions.

Proposition. If X and Y are independent, then $\sigma(X,Y) = 0$. (Note that the converse is not true.)

Proposition (Linear Regression: Estimation of the Regression Slope)**.** Let X and Y be random variables with means $\mu_X = E(X)$ and $\mu_Y = E(Y)$, variance $\sigma^2(X)$, and covariance $\sigma(X,Y)$. In modeling the relationship between X and Y by the linear model

$$Y = \mu_Y + b\,(X - \mu_X),$$

the least-squares criterion leads to the unique optimal slope

$$b = \frac{\sigma(X,Y)}{\sigma^2(X)}.$$

In other words, the line $Y = \mu_Y + b\,(X-\mu_X)$ minimizes the expected squared error

$$K(b) = E\Big[(Y - \mu_Y - b\,(X-\mu_X))^2\Big].$$

Moreover, when applied to sample data $\{(x_i, y_i)\}_{i=1}^N$ with sample means

$$\hat{\mu}_X = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \hat{\mu}_Y = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

the regression slope is estimated by

$$\hat{b} = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)}{\sum_{i=1}^{N}(x_i - \hat{\mu}_X)^2}.$$

5 Measure Theoretic Integration

Definition (Outer Measure)**.** The *outer measure* of a subset $A \subset \mathbb{R}^n$ is defined by

$$\lambda^*(A) = \inf \sum_{k=1}^{\infty} \mathrm{vol}(C_k),$$

where the infimum is taken over all countable collections of sets $C_1, C_2, \cdots \subset \mathbb{R}^n$ satisfying

$$A \subset \bigcup_{k=1}^{\infty} C_k.$$

Definition (λ^*-Measurable Sets)**.** A set $A \subset \mathbb{R}^n$ is called λ^*-*measurable* if for every $B \subset \mathbb{R}^n$, we have

$$\lambda^*(B) = \lambda^*(A \cap B) + \lambda^*(A^c \cap B).$$

Remark. This measurability condition turns out to be equivalent to an *inner-measure equals outer-measure* type equality, similar to the criterion for Riemann integrability.

Proposition. Any countable set has outer measure zero.

Proposition. Any set of outer measure zero is measurable.

Proposition. The collection $\mathcal{F}$ of all λ^*-measurable sets is a σ-algebra, and the function λ^* restricted to $\mathcal{F}$ defines a measure λ.

Definition. The measure λ is called the *Lebesgue measure*.

Proposition. If C is a cube in $\mathbb{R}^n$, then $C \in \mathcal{F}$, and its Lebesgue measure satisfies

$$\lambda(C) = \mathrm{vol}(C).$$

Definition (Indicator Function)**.** Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and fix a set $A \in \mathcal{F}$. The *indicator function* of A is the function $\mathbb{1}_A : \Omega \to \mathbb{R}$ defined by

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{if } \omega \in A^c. \end{cases}$$

Proposition. For any indicator function $\mathbb{1}_A$, we define its integral as

$$\int_\Omega \mathbb{1}_A(\omega) d\mu(\omega) = \mu(A).$$

Definition (Simple Function)**.** A *simple function* is a function of the form

$$f = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i},$$

where $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$.

Proposition. For any simple function f, its integral is defined as

$$\int_\Omega f(\omega) d\mu(\omega) = \sum_{i=1}^{n} a_i \cdot \mu(A_i).$$

Remark. Any random variable $X : \Omega \to \mathbb{R}$ with a finite range of values is a simple function. The above integral formula agrees with the expected value formula:

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i \cdot \mu(A_i),$$

where $A_i = X^{-1}(\{a_i\})$.

Definition (Decomposition of a Measurable Function)**.** Let $f : \Omega \to \mathbb{R}$ be a measurable function. We can write

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

where $f^+, f^- : \Omega \to \mathbb{R}$ are measurable and nonnegative, given by

$$f^+ = \max(f, 0), \quad f^- = \max(-f, 0).$$

Theorem. Let $f : \Omega \to \mathbb{R}$ be a nonnegative function. Then f is measurable if and only if it is the limit of an increasing sequence of nonnegative simple functions.

Definition (Lebesgue Integral for Nonnegative Functions)**.** Let $(f_n)_{n\geq 1}$ be an increasing sequence of non-negative simple functions such that $f_n \to f$ pointwise. Then the integral of f is defined as

$$\int_\Omega f(\omega)d\mu(\omega) = \lim_{n\to\infty} \int_\Omega f_n(\omega)d\mu(\omega).$$

Definition (Lebesgue Integral for General Functions)**.** For an arbitrary measurable function $f : \Omega \to \mathbb{R}$, write $f = f^+ - f^-$. The integral of f is then defined as

$$\int_\Omega f(\omega)d\mu(\omega) = \int_\Omega f^+(\omega)d\mu(\omega) - \int_\Omega f^-(\omega)d\mu(\omega),$$

provided that at least one of the integrals on the right-hand side is finite.

Theorem. Every Riemann integrable function is measurable, and its Lebesgue integral equals its Riemann integral.

Proposition. The collection $\mathcal{F}$ of all λ^*-measurable sets is a σ-algebra, and the function λ^* restricted to $\mathcal{F}$ defines a measure λ.

Proposition. Every set $B \subset \mathbb{R}^n$ with $\lambda^*(B) = 0$ is measurable.

Definition. The Borel σ-algebra on $\mathbb{R}^n$, denoted $\mathcal{B}(\mathbb{R}^n)$, is the σ-algebra generated by all open sets $U \subset \mathbb{R}^n$.

Proposition. Every open subset of $\mathbb{R}^n$ is λ-measurable.

Corollary. We have the inclusion

$$\mathcal{B}(\mathbb{R}^n) \subset \mathcal{F},$$

so the measure λ restricts to a measure on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$.

Remark. It turns out that $\mathcal{F}$ is the *completion* of $\mathcal{B}(\mathbb{R}^n)$. This means that $\mathcal{F}$ includes all sets in $\mathcal{B}(\mathbb{R}^n)$ along with additional subsets of null sets. While this makes $\mathcal{F}$ larger and more inclusive, it can also make the measurability of functions mapping into $\mathbb{R}^n$ more challenging.

6 General Random Variables and Cumulative Distribution Functions

Definition (Random Variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be the Borel measurable space. A **random variable (R.V.)** is a measurable function

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})).$$

Corollary. A random variable X defines a probability measure P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by

$$P_X(B) = P(X^{-1}(B)), \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

Proposition. The Borel σ-algebra $\mathcal{B}(\mathbb{R})$ is generated by the collection of half-rays $(-\infty, x]$ for $x \in \mathbb{R}$.

Corollary. A function X is a random variable if and only if for all $x \in \mathbb{R}$,

$$\{X \leq x\} := X^{-1}(-\infty, x] \in \mathcal{F}.$$

Remark. This condition is the reason for the definition of a random variable found in standard probability textbooks.

Definition (Cumulative Distribution Function)**.** The **cumulative distribution function (CDF)** of a random variable X is the function $F_X : \mathbb{R} \to [0, 1]$ given by

$$F_X(x) = P_X((-\infty, x]) = P(\{X \leq x\}).$$

Lemma. Let $(A_n)_{n\geq 1}$ be an increasing sequence of events, i.e., $A_1 \subset A_2 \subset \dots$, or let $(B_n)_{n\geq 1}$ be a decreasing sequence of events, i.e., $B_1 \supset B_2 \supset \dots$. Then

$$\lim_{n\to\infty} P(A_n) = P\left(\bigcup_{i=1}^{\infty} A_i\right),$$

$$\lim_{n\to\infty} P(B_n) = P\left(\bigcap_{i=1}^{\infty} B_i\right).$$

Corollary. The cumulative distribution function satisfies

$$\lim_{x\to\infty} F_X(x) = 1, \quad \lim_{x\to-\infty} F_X(x) = 0.$$

Corollary. The function $F_X(x)$ is right-continuous, but not necessarily continuous.

7 Absolute Continuity and Continuous Random Variables

Let $(E, \mathcal{E})$ be a measurable space, μ a measure on $(E, \mathcal{E})$, and let $f : E \to \mathbb{R}$ be a positive, measurable function. Then:

Proposition. The function $\nu_f : \mathcal{E} \to \mathbb{R}$ given by

$$\nu_f(A) = \int_A f \, d\mu = \int_E 1_A \cdot f \, d\mu$$

defines a measure on $(E, \mathcal{E})$.

Definition (Absolute Continuity)**.** Let $(E, \mathcal{E})$ be a measurable space, and let μ and ν be measures on $(E, \mathcal{E})$. We say that ν is **absolutely continuous** with respect to μ, written $\nu \ll \mu$, if

$$\mu(A) = 0 \quad \text{implies} \quad \nu(A) = 0, \quad \forall A \in \mathcal{E}.$$

Proposition. The measure ν_f defined above is absolutely continuous with respect to μ.

Theorem (Radon-Nikodym Theorem)**.** Let ν be a measure that is absolutely continuous with respect to μ, and assume that μ is σ-finite. Then there exists a measurable function $f : E \to \mathbb{R}$ such that

$$\nu(A) = \int_A f \, d\mu, \quad \forall A \in \mathcal{E}.$$

The function f is called the **Radon-Nikodym derivative** and is denoted by

$$\frac{d\nu}{d\mu}.$$

Remark. The Radon-Nikodym derivative $\frac{d\nu}{d\mu}$ is unique up to redefinition on a set of μ-measure zero.

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Definition (Continuous Random Variable)**.** A random variable $X : \Omega \to \mathbb{R}$ is called **continuous** if the induced measure P_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is absolutely continuous with respect to the Lebesgue measure λ.

Equivalently, there exists a measurable function $f_X : \mathbb{R} \to \mathbb{R}$ such that for all $A \in \mathcal{B}(\mathbb{R})$,

$$P_X(A) = \int_A f_X \, d\lambda = \int_A f_X(x) \, dx.$$

Definition (Probability Density Function)**.** The function $f_X : \mathbb{R} \to \mathbb{R}$ is called the **probability density function** (PDF) of X.

Remark. It suffices to check the above condition for $A = (-\infty, x]$, i.e.,

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u) \, du.$$

Remark. By the Fundamental Theorem of Calculus, this implies that

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

Remark. Conversely, given any positive, measurable function $f : \mathbb{R} \to \mathbb{R}$ such that

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

we can define F_X as above, or equivalently define the induced measure P_X.

Theorem (Jacobian Formula)**.** Let X and Y be jointly continuous with joint density function $f_{X,Y}$, and let

$$D = \{(x, y) : f_{X,Y}(x, y) > 0\}.$$

If the mapping T given by

$$T(x, y) = (u(x, y), v(x, y))$$

is a bijection from D to the set $S \subseteq \mathbb{R}^2$, then (subject to the previous conditions) the pair $(U, V) = (u(X, Y), v(X, Y))$ is jointly continuous with joint density function

$$f_{U,V}(u, v) = \begin{cases} f_{X,Y}(x(u, v), y(u, v))\, |J(u, v)|\,, & \text{if } (u, v) \in S, \\ 0, & \text{otherwise.} \end{cases}$$

Proof. You should not worry overmuch about the details of this argument. Suppose that $A \subseteq D$ and $T(A) = B$. Since $T : D \to S$ is a bijection,

$$\mathbb{P}((U, V) \in B) = \mathbb{P}((X, Y) \in A).$$

However,

$$\mathbb{P}((X, Y) \in A) = \iint_A f_{X,Y}(x, y)\, dx\, dy \quad \text{by Theorem 6.22.}$$

Using the change of variables formula,

$$\mathbb{P}((X, Y) \in A) = \iint_B f_{X,Y}(x(u, v), y(u, v))\, |J(u, v)|\, du\, dv.$$

By (6.52), we conclude

$$\mathbb{P}((U, V) \in B) = \iint_B f_{U,V}(u, v)\, du\, dv.$$

$\square$

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable. Given a measurable function $h : \mathbb{R} \to \mathbb{R}$, we define another random variable $Y = h \circ X : \Omega \to \mathbb{R}$.

By the pushforward measure transformation, we have:

$$P_Y := Y_* P = h_* X_* P = h_* P_X,$$

which implies that for any measurable set $A \in \mathcal{B}(\mathbb{R})$,

$$P_Y(A) = P(Y^{-1}(A)) = P(X^{-1}(h^{-1}(A))) = P_X(h^{-1}(A)).$$

Proposition. Let $h : \mathbb{R} \to \mathbb{R}$ be a differentiable and strictly increasing function on the support of X, denoted $V_X \subset \mathbb{R}$. Then $Y = h(X) = h \circ X$ is a continuous random variable with probability density function given by:

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

Remark. This result is compatible with the general transformation formula for probability measures. Specifically, we have:

$$P_X(h^{-1}(A)) = \int_{x \in h^{-1}(A)} f_X(x)\, dx$$

which, via the change of variables $y = h(x)$, transforms into:

$$P_Y(A) = \int_{y \in A} f_X(h^{-1}(y)) \frac{d}{dy} h^{-1}(y)\, dy.$$

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable.

Definition (Expected Value)**.** The expected value of X, denoted $\mathbb{E}[X]$, is defined as

$$\mathbb{E}[X] = \int_\Omega X\, dP = \int_{\omega \in \Omega} X(\omega) dP(\omega),$$

where integration is defined in the measure-theoretic sense.

Remark. For discrete random variables, which take countable values, the expected value reduces to:

$$\mathbb{E}[X] = \sum_{x \in V_X} x \cdot P(X = x) = \sum_{x \in V_X} x \cdot f_X(x).$$

Remark. For continuous random variables, the expected value is given by:

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

Proposition (Expectation of a Function of a Random Variable)**.** Let $X : \Omega \to \mathbb{R}$ be a random variable, and let $h : \mathbb{R} \to \mathbb{R}$ be a measurable function. Define $Y = h(X)$. Then the expectation of Y is given by:

$$\mathbb{E}[Y] = \mathbb{E}[h(X)] = \int_\mathbb{R} h(x) f_X(x)\, dx.$$

Remark. This is the continuous analogue of the discrete case:

$$\mathbb{E}[h(X)] = \sum_{x \in V_X} h(x) f_X(x).$$

Definition (Poisson Distribution)**.** Consider a process where events occur continuously in time, with the number of subintervals n tending to infinity while the probability of an event occurring in any subinterval tends to zero. The probability mass function for the total number of events is given by:

$$P(X = k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Taking the limit as $n \to \infty$, we obtain:

$$\frac{n!}{(n-k)!} \frac{1}{n^k} \to 1, \quad \text{and} \quad \left(1 - \frac{\lambda}{n}\right)^{n-k} \to e^{-\lambda}.$$

Thus, we define the Poisson distribution as:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k \in \mathbb{Z}_{\geq 0}.$$

Formally, let $V_X = \mathbb{Z}_{\geq 0}$ and define $f_X : V_X \to \mathbb{R}_{\geq 0}$ by:

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

The resulting measure P_X satisfies:

$$\mathbb{E}[X] = \lambda, \quad \text{and} \quad \mathrm{Var}(X) = \lambda.$$

Definition (Stationary Poisson Counting Process)**.** A *stationary Poisson counting process* of rate λ is a stochastic process $\{N_t : \Omega \to \mathbb{Z}_{\geq 0}\}_{t \in \mathbb{R}_{\geq 0}}$ satisfying the following conditions:

- **(Normalization)** $N_0 = 0$.
- **(Increasing)** $N_t - N_s$ takes non-negative values for $s < t$.
- **(Independent increments)** The random variables $N_t - N_{s_1}$ and $N_{s_1} - N_{s_0}$ are independent for any $0 \leq s_0 < s_1 < t$ (similarly, increments are mutually independent for multiple disjoint intervals).
- **(Poisson property)** $N_t - N_s$ is Poisson distributed with mean $\lambda(t - s)$.

Proposition (Arrival Times and the Exponential Distribution)**.** Let $\{N_t : \Omega \to \mathbb{Z}_{\geq 0}\}_{t \in \mathbb{R}_{\geq 0}}$ be a Poisson counting process of rate λ. Each random variable N_t follows a Poisson distribution with mean λt, representing the number of arrivals occurring in the interval $[0, t]$.

Consider the random variable $X_1 : \Omega \to \mathbb{R}_{\geq 0}$ representing the time of the first arrival, defined as:

$$\{X_1 \leq t\} := \{N_t \geq 1\}.$$

That is, for an outcome $\omega \in \Omega$, the first arrival $X_1(\omega)$ occurs at time $\leq t$ if and only if the number of arrivals $N_t(\omega)$ up to time t is at least 1.

The cumulative distribution function of X is given by:

$$F_X(x) = P(X \leq x) = P(N_x \geq 1) = 1 - P(N_x = 0).$$

Since N_x follows a Poisson distribution with mean λx, we have:

$$P(N_x = 0) = e^{-\lambda x}.$$

Thus,

$$F_X(x) = 1 - e^{-\lambda x}.$$

Differentiating, we obtain the probability density function:

$$f_X(x) = \frac{d}{dx} F_X(x) = \lambda e^{-\lambda x}.$$

Proposition (Basic Properties of the Exponential Distribution)**.** Let $\theta = \frac{1}{\lambda}$, and let X be a continuous random variable with support $V_X = \mathbb{R}_{\geq 0}$ and probability density function:

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0.$$

The exponential distribution satisfies the following properties:

1. **Expectation:** The expected value of X is given by:

$$\mathbb{E}[X] = \int_0^\infty x f_X(x)\, dx.$$

   Substituting $f_X(x)$ and using integration by parts, we obtain:

$$\mathbb{E}[X] = \theta.$$

2. **Variance:** The variance of X is given by:

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$

   Computing $\mathbb{E}[X^2]$ via integration by parts and simplifying, we find:

$$\mathrm{Var}(X) = \theta^2.$$

3. **Memoryless Property:** The exponential distribution satisfies:

$$P(X > t + s \mid X > s) = P(X > t).$$

   By definition of conditional probability,

$$P(X > t + s \mid X > s) = \frac{P(X > t + s)}{P(X > s)}.$$

   Since $P(X > x) = \int_x^\infty f_X(u)\, du = e^{-x/\theta}$, we compute:

$$P(X > t + s \mid X > s) = \frac{e^{-(t+s)/\theta}}{e^{-s/\theta}} = e^{-t/\theta} = P(X > t).$$

   This property encodes a type of *sunk-cost fallacy*: given that one has already waited s minutes with no occurrences, the probability of waiting an additional t minutes is the same as if one had just started waiting.

Experiment Type	Fixed Period	Until 1st Success	Until αth Success
Coin Flipping	Binomial	Geometric	Negative Binomial
Continuous Arrivals	Poisson	Exponential	Gamma

Definition (Gamma Distribution and Gamma Function)**.** The Gamma distribution generalizes the exponential distribution and arises in the context of waiting times until the αth event in a Poisson process.

A *Gamma-distributed* continuous random variable W with shape parameter $\alpha > 0$ and scale parameter $\theta > 0$ has the probability density function:

$$f_W(w) = \frac{w^{\alpha-1}}{\theta^\alpha \Gamma(\alpha)} e^{-w/\theta}, \quad w > 0.$$

When $\alpha \in \mathbb{Z}_{\geq 1}$, this simplifies to:

$$f_W(w) = \frac{w^{\alpha-1}}{\theta^\alpha (\alpha-1)!} e^{-w/\theta}.$$

The Gamma function $\Gamma : \mathbb{R}_{>0} \to \mathbb{R}$ is defined as:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} \, dy.$$

This function extends the factorial to non-integer values, satisfying:

$$\Gamma(1) = 1, \quad \text{and} \quad \Gamma(z+1) = z\Gamma(z).$$

As a consequence, for integer values $\alpha \in \mathbb{Z}_{\geq 1}$, we recover:

$$\Gamma(\alpha) = (\alpha - 1)!.$$

The Gamma distribution has expectation and variance given by:

$$\mathbb{E}[W] = \alpha\theta, \quad \mathrm{Var}(W) = \alpha\theta^2.$$

Definition (Chi-Square Distribution)**.** The *chi-square distribution*, denoted χ_r^2, with r degrees of freedom, is a special case of the Gamma distribution with parameters:

$$\alpha = \frac{r}{2}, \quad \theta = 2.$$

Explicitly, the probability density function is given by:

$$f_X(x) = \frac{1}{\Gamma(r/2) 2^{r/2}} x^{(r/2)-1} e^{-x/2}, \quad x > 0.$$

By the properties of the Gamma distribution, the expectation and variance of X are:

$$\mathbb{E}[X] = \alpha\theta = r, \quad \mathrm{Var}(X) = \alpha\theta^2 = 2r.$$

While the chi-square distribution may seem abstract in this formulation, it plays a fundamental role in statistical theory. It describes the distribution of the sum of squares of r independent standard normal random variables, making it essential in hypothesis testing, confidence intervals, and variance estimation.

8 Generating Functions

Definition (Generating Function)**.** A **generating function** of a random variable X taking values in $\mathbb{Z}_{\geq 0}$ is defined as:

$$G_X(x) = \sum_{k=0}^{\infty} \mathbb{P}(X = k)x^k.$$

(Good theorem: Power series converges in radius R).

Proposition. Let X and Y be independent random variables. Then their generating functions satisfy:

$$G_{X+Y}(x) = G_X(x)G_Y(x).$$

Proof. Consider the coefficient of x^n on the left-hand side:

$$\text{LHS}[x^n] = \mathbb{P}(X + Y = n).$$

By the law of total probability, we sum over all possible values of X:

$$\sum_k \mathbb{P}(X = k, Y = n - k).$$

Since X and Y are independent, this factorizes as:

$$\sum_k \mathbb{P}(X = k)\mathbb{P}(Y = n - k).$$

This is precisely the coefficient of x^n in the right-hand side:

$$\text{RHS}[x^n].$$

Hence, the two generating functions satisfy the claimed equality. □

Definition (Convolution)**.** A convolution of two sequences $(a_n)_{n=0}^{\infty}$ and $(b_n)_{n=0}^{\infty}$ is the sequence defined by:

$$(a * b)_n = \sum_{k \leq n} a_k b_{n-k} = \sum_{k=0}^{n} a_k b_{n-k}.$$

Proposition. The probability mass function of the sum of two independent discrete random variables is given by the convolution:

$$(\mathbb{P}(X + Y = n))_{n=0}^{\infty} = \mathbb{P}(X = n) * \mathbb{P}(Y = n).$$

Definition. For functions $f, g : \mathbb{R} \to \mathbb{R}$, the convolution is defined as:

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t)g(x - t)\, dt.$$

Proposition. If X and Y are independent random variables with density functions $f_X(x)$ and $g_X(x)$, then their sum $Z = X + Y$ is a continuous random variable with density given by the convolution:

$$f * g.$$

Proof. We compute the probability:

$$\mathbb{P}(Z \leq t) = \mathbb{P}(X + Y \leq t).$$

By integrating over the joint density,

$$\mathbb{P}(Z \leq t) = \int_{-\infty}^{\infty} \int_{-\infty}^{t-x} f(x)g(y)\,dy\,dx.$$

Using the transformation $Z = X + Y$, $Y = Z - X$, we get:

$$\mathbb{P}(Z \leq t) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)g(z - x)\,dx\,dz.$$

Since the Jacobian determinant of the transformation is 1, we obtain the convolution formula for the density:

$$f_Z(z) = \int_{-\infty}^{\infty} f(x)g(z - x)\,dx.$$

□

Proposition. If X and Y are independent continuous random variables, then their density satisfies:

$$f_{X+Y}(t) = f_X * f_Y(t).$$

Proof. By the definition of convolution,

$$g * f(t) = f * g(t),$$

so we compute:

$$\int_{-\infty}^{\infty} g(x)f(t - x)\,dx.$$

Making the substitution $y = t - x$, so that $dy = -dx$, we rewrite the integral as:

$$\int_{-\infty}^{\infty} f(y)g(t - y)\,dy.$$

This confirms the convolution formula for independent continuous random variables.

Moreover, this definition is well-defined in the sense that convolution satisfies associativity:

$$(f * g) * h = f * (g * h).$$

□

Theorem (Central Limit Theorem)**.** Let $X_1, X_2, \ldots$ be independent and identically distributed random variables with mean 0 and variance σ^2. Then,

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as } n \to \infty.$$

Theorem (Nice Properties of Normal Distribution).

$$\mathcal{N}(0,\sigma^2) = \mathcal{N}(0,\sigma_1^2) * \mathcal{N}(0,\sigma_2^2) \neq \mathcal{N}(0,\sigma_2^2)$$

for $\sigma_1^2 + \sigma_2^2 = \sigma^2$.

Theorem (Derivation of the Normal Distribution). 1. $\rho(x,y) = f(x)f(y)$

2. $\rho(x,y) = g\left(\sqrt{x^2+y^2}\right)$
   Thus,

$$f(0)\cdot f(0) = g(0) \Rightarrow g(x) = \frac{f(x)}{f(0)}$$

$$g\left(\sqrt{x^2+y^2}\right) = g(x)g(y)c^2$$

Let $h = \ln(g)\cdot c^2$,

$$h\left(\sqrt{x^2+y^2}\right) = h(x) + h(y)$$

which implies that

$$j(x^2+y^2) = j(x^2) + j(y^2)$$

for some function $j(x)$, leading to

$$j(x) = \alpha x, \quad h(x) = cx^2$$

$$g(x) = e^{cx^2}\cdot c$$

$$\Rightarrow f(x) = e^{cx^2}\cdot c$$

which leads to the well-known Gaussian density function.

Proposition. If g is an increasing function and X is a continuous random variable, then

$$\rho_{g(X)}(g(t)) = \rho_X(t)\cdot\left|g'(t)\right|.$$

Proof. The cumulative distribution function transformation gives:

$$F_{g(X)}(g(t)) = \mathbb{P}(X < t) = F_X(t).$$

Differentiating both sides with respect to t,

$$\rho_{g(X)}(g(t))\cdot g'(t) = \rho_X(t).$$

Rearranging, we obtain:

$$\rho_{g(X)}(g(t)) = \frac{\rho_X(t)}{g'(t)}.$$

□

Proposition. If g is a decreasing function and X is a continuous random variable, then

$$\rho_{g(X)}(g(t)) = \frac{\rho_X(t)}{-g'(t)}.$$

Proof. The cumulative distribution function transformation gives:

$$F_{g(X)}(g(t)) = \mathbb{P}(g(X) \leq g(t)) = 1 - \mathbb{P}(X < t) = 1 - F_X(t).$$

Differentiating both sides with respect to t,

$$\rho_{g(X)}(g(t)) \cdot g'(t) = \rho_X(t).$$

Rearranging, we obtain:

$$\rho_{g(X)}(g(t)) = \frac{\rho_X(t)}{-g'(t)}.$$

$\square$

Theorem (Change of Variables)**.** Suppose $g \in C^1$ is a one-to-one function $g : U \to V$, where $U, V \subseteq \mathbb{R}^n$, and let X be a random variable taking values in U. Then, the density of $g(X)$ is given by

$$\rho_{g(X)}(g(\vec{t})) = \frac{\rho_X(\vec{t})}{|J_g(\vec{t})|}.$$

Proposition. If X and Y are independent continuous random variables with $Y > 0$, then their ratio has the density

$$\rho_{X/Y}(t) = \int_{-\infty}^{+\infty} f(x) g\left(\frac{t}{x}\right) \frac{dx}{x}.$$

Proof. We begin by computing the cumulative distribution function:

$$F_{X/Y}(t) = \mathbb{P}(X/Y < t).$$

Rewriting this probability,

$$F_{X/Y}(t) = \mathbb{P}(X, Y \in A_t),$$

where A_t is the region satisfying $X < tY$. Using the joint density $\rho_{X,Y}(x, y)$, we express this as the integral

$$F_{X/Y}(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{tx} \rho_{X,Y}(u, v)\, du\, dv.$$

Using the transformation $u = x, v = xy$, the integral changes to

$$F_{X/Y}(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{t} \rho_{X,Y}(u, v/u) \frac{du}{u}\, dv.$$

Differentiating both sides with respect to t, we obtain

$$\rho_{X/Y}(t) = \int_{-\infty}^{+\infty} \rho_{X,Y}(u, t/u) \frac{du}{u}.$$

Since X and Y are independent, we factorize the joint density:

$$\rho_{X/Y}(t) = \int_{-\infty}^{+\infty} \rho_X(u) \rho_Y(t/u) \frac{du}{u}.$$

$\square$

Definition (Normal Distribution)**.** Fix $\mu \in \mathbb{R}$ and $\sigma > 0$. We say a continuous random variable X follows a normal distribution with mean μ and variance σ^2, denoted $X \sim N(\mu, \sigma^2)$, if

$$V_X = \mathbb{R}, \quad f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We will always take as given that

$$\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-z^2/2}\, dz = 1,$$

which is proved by squaring and computing the double integral in radial coordinates.

It is easy to check using the change of variables formula that

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

in general, and similarly that

$$\mathbb{E}[X] = \mu, \quad \mathrm{Var}(X) = \sigma^2,$$

justifying the names.

Mnemonic: If $X \sim N(\mu, \sigma^2)$, then the standardized random variable

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

follows the standard normal distribution, $Y \sim N(0, 1)$.

Joint Continuity and Bivariate distributions

Definition (Jointly Continuous Random Variables)**.** Random variables X and Y are **jointly continuous** if $P(X, Y)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. Equivalently, there exists a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_{\geq 0}$ such that for any measurable set A,

$$P(X, Y)(A) = \int_A f_{X,Y}(x, y)\, dx\, dy.$$

Remark. Equivalently, it suffices to check that the joint cumulative distribution function (CDF) satisfies

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

Definition (Marginal Densities)**.** The **marginal densities** of a jointly continuous random variable (X, Y) with density $f_{X,Y}$ are given by:

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy, \quad f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dx.$$

Proposition. Let X, Y be continuous random variables with densities f_X and f_Y. Recall that X and Y are independent if and only if their joint cumulative distribution function satisfies

$$F_{X,Y}(x, y) = F_X(x) F_Y(y).$$

Proposition. If X and Y are jointly continuous, then X and Y are independent if and only if

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

almost everywhere.

Remark. A stronger statement holds: If there exist functions $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ such that

$$f_{X,Y}(x,y) = h(x)g(y),$$

then we can conclude that $h = f_X$, $g = f_Y$, and X, Y are independent.

Definition (Joint CDF)**.** The **joint cumulative distribution function (CDF)** of random variables X and Y, denoted $F_{X,Y}$, is the function $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ defined by:

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y).$$

Equivalently,

$$F_{X,Y}(x,y) = P_{(X,Y)}((-\infty, x] \times (-\infty, y]).$$

Remark. The Borel sigma-algebra on $\mathbb{R}^2$, denoted $\mathcal{B}(\mathbb{R}^2)$, is generated by sets of the form $(-\infty, x] \times (-\infty, y]$ for $x, y \in \mathbb{R}$. Thus, the function $F_{X,Y}$ faithfully encodes the probability measure $P_{(X,Y)}$, as we saw with the univariate CDF F_X and the probability measure P_X.

Proposition. The joint CDF $F_{X,Y}$ satisfies the following properties:

1. $F_{X,Y}$ is non-decreasing in both x and y.
2. $F_{X,Y}$ is right-continuous in both x and y.
3. $\lim_{x,y \to -\infty} F_{X,Y}(x,y) = 0$ and $\lim_{x,y \to +\infty} F_{X,Y}(x,y) = 1$.
4. $\lim_{x \to +\infty} F_{X,Y}(x,y) = F_Y(y)$ and $\lim_{y \to +\infty} F_{X,Y}(x,y) = F_X(x)$.

Proposition. The random variables X and Y are independent if and only if

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Definition (Conditional Density Function)**.** The **conditional density function** of X given Y is defined as:

$$f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

for any fixed $y \in \mathbb{R}$ such that $f_Y(y) > 0$.

Proposition. The function $f_{X|Y}(x|y)$ satisfies the properties of a probability density function (pdf) and defines a conditional probability measure $P_{X|Y=y}$ on $\mathbb{R}$.

Proposition. Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $X, Y : \Omega \to \mathbb{R}$ be random variables. Suppose that (X, Y) are jointly continuous with density $f_{X,Y}$. Let $h : \mathbb{R}^2 \to \mathbb{R}$ be a measurable function, and define $Z = h(X, Y)$. Then, the expected value of Z is given by:

$$E(Z) = E(h(X, Y)) = \int_{\mathbb{R}^2} h(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

Proposition (Measure-Theoretic Change of Variables)**.** Let $(E_1, \mathcal{E}_1, \mu)$ be a measure space, and let $(E_2, \mathcal{E}_2)$ be another measurable space. Suppose $g : E_1 \to E_2$ is measurable. Then, for any measurable function $h : E_2 \to \mathbb{R}$ such that the integrals converge, we have:

$$\int_{e \in E_1} (h \circ g)(e)\, d\mu(e) = \int_{x \in E_2} h(x)\, d(g_*\mu)(x),$$

where $g_*\mu$ denotes the pushforward measure $\mu \circ g^{-1}$.

Applying this to probability spaces, take $E_1 = \Omega$, $E_2 = \mathbb{R}^2$, $g = (X, Y)$, and $\mu = P$, so that the pushforward measure $g_*\mu$ corresponds to $P_{(X,Y)}$. This yields the result:

$$E(h(X, Y)) = \int_{\mathbb{R}^2} h(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

Definition (Standard Bivariate Normal Distribution)**.** Let X and Y be jointly continuous random variables with support $V_{X,Y} = \mathbb{R}^2$ and joint density function given by:

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right).$$

Proposition. The function $f_{X,Y}$ integrates to 1, thereby defining a probability measure.

To compute the marginal density $f_X(x)$ of X, we use the identity:

$$x^2 - 2\rho xy + y^2 = x^2 - \rho^2 x^2 + \rho^2 x^2 - 2\rho xy + y^2 = x^2(1-\rho^2) + (y - \rho x)^2.$$

Thus, integrating out y gives:

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)\, dy.$$

Substituting the density function:

$$\int_{\mathbb{R}} \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2(1-\rho^2)+(y-\rho x)^2}{2(1-\rho^2)}}\, dy.$$

Recognizing the inner integral as the density of $N(\rho x, 1-\rho^2)$, which integrates to 1, we obtain:

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

Thus, $f_X(x)$ follows $N(0, 1)$, and by symmetry, so does $f_Y(y)$.

Definition (General Bivariate Normal Distribution)**.** Fix $\mu_X, \mu_Y \in \mathbb{R}$ and $\sigma_X, \sigma_Y > 0$. The joint density function of the bivariate normal distribution is given by:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right)\right).$$

Proposition. The marginal density $f_X(x)$ is computed by a change of variables. Setting $\tilde{y} = \frac{y-\mu_Y}{\sigma_Y}$, we obtain:

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right) \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(\tilde{y}-\rho x)^2}{2(1-\rho^2)}\right) d\tilde{y}.$$

Recognizing the inner integral as the density function of $N(\rho x, 1-\rho^2)$, which integrates to 1, we conclude:

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right).$$

Thus, $X \sim N(\mu_X, \sigma_X^2)$, and by symmetry, $Y \sim N(\mu_Y, \sigma_Y^2)$.

Similarly, the conditional density $f_{X|Y}(x|y)$ is normally distributed as:

$$X|Y \sim N\left(\rho\tilde{y}, 1-\rho^2\right).$$

Finally, we can verify that the covariance satisfies:

$$\sigma(X,Y) = \rho\sigma_X\sigma_Y.$$

Definition. A sequence $\{X_n : \Omega \to \mathbb{R}\}$ converges **almost surely** to $X : \Omega \to \mathbb{R}$ if

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

This is a rather strong notion of convergence. Another strong notion, which is neither strictly stronger nor weaker, is:

Definition. $X_n \to X$ in **mean square** if

$$\lim_{n\to\infty} E\left((X - X_n)^2\right) = 0.$$

Both of the preceding are strictly stronger than the following:

Definition. $X_n \to X$ in **probability** if

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

for any $\varepsilon > 0$. This is also strictly stronger than:

Definition. $X_n \to X$ in **distribution** if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for every point $x \in \mathbb{R}$ at which F_X is continuous.

Theorem (Law of Large Numbers)**.** The law of large numbers encodes the idea that if we have samples $x_1, \ldots, x_n$ from some abstract distribution X, the sample mean

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \to \mu = E(X) \quad \text{as } n \to \infty.$$

This is an abstract theorem in the following sense:

Let $X_1, \ldots, X_n$ be independent, identically distributed random variables with finite mean μ and variance σ^2. Then the random variable

$$X_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

satisfies $X_n \to \mu$ in mean square.

Thus, as $n \to \infty$, this random variable converges (in mean square) to the deterministic (i.e., constant) random variable taking the value μ.

Proposition. In a sense, we can interpret this as a drawback of the sample mean: In the limit where the number of samples $n \to \infty$, the random variable X_n contains only the information of the mean μ. Heuristically, this is because the difference

$$\lim_{n\to\infty} X_n - \mu = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n - n\mu}{n} = 0$$

is dominated by the denominator n. We can also see that the distribution is becoming very peaked because the variance $\to 0$:

$$\sigma^2(X_n) = \sigma^2\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}\sigma^2(X_1 + \cdots + X_n) = \frac{\sigma^2(X)}{n}.$$

Again, the factor $\frac{1}{n}$ from the definition of the mean dominates! The random variable $X_1 + \cdots + X_n$ doesn't converge to anything, as its variance

$$\sigma^2(X_1 + \cdots + X_n) = n\sigma^2(X)$$

goes to infinity as $n \to \infty$.

In summary, the denominator n dominates in the limit

$$\lim_{n\to\infty} X_n - \mu = \lim_{n\to\infty} \frac{X_1 + \cdots + X_n - n\mu}{n} = 0,$$

while $X_1 + \cdots + X_n - n\mu$ has variance $n\sigma^2(X) \to \infty$ as $n \to \infty$.

Motivated by this, we consider the distribution of the stable ratio

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}},$$

which satisfies

$$\sigma^2(Y_n) = \sigma^2(X).$$

Taking the limit $n \to \infty$ while holding the variance constant, we obtain the following theorem.

Theorem (Central Limit Theorem)**.** Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed random variables with finite mean μ and variance σ^2. Then the sequence of random variables

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}}$$

converges in distribution to a normal distribution with mean 0 and variance $\sigma^2(X)$:

$$Y_n \xrightarrow{d} N(0, \sigma^2).$$

Combinatorics Applications

Proposition (The Probabilistic Method)**.** We want to construct an object with a desirable property, but direct construction may be difficult. Instead, we take a random such object. If none of these objects satisfy the desired property, then

$$\mathbb{P}(\text{random object is nice} = 0).$$

However, if

$$\mathbb{P}(\text{random object is nice} > 0),$$

then at least one such object must exist.

Definition (Ramsey Number)**.** Let K_n denote the complete graph on n vertices, where every pair of vertices is connected by an edge. Consider an edge coloring where each edge is colored either red or blue.

We define the *Ramsey number* $R(k, \ell)$ as the smallest integer n such that in any red-blue edge coloring of K_n, there exists either:

- a subset of k vertices where all edges between them are red, or
- a subset of ℓ vertices where all edges between them are blue.

Theorem. The Ramsey number $R(k, \ell)$ is always finite.

Example. We have the specific values:

$$R(3,3) = 6, \quad R(5,5) \text{ (unknown)}.$$

Theorem. For all $k \geq 3$, the Ramsey number satisfies the bound:

$$R(k,k) > 2^{k/2} - 1.$$

Definition (Tournament)**.** A *tournament* is a directed complete graph, where every player (vertex) plays every other player, and an arrow points to the winner.

Remark. Tournaments can be used to produce rankings, starting from the lowest-ranked player, by hitting every vertex exactly once in a path.

Definition (Hamiltonian Path)**.** A *Hamiltonian path* is a path that visits each vertex exactly once.

Theorem. Every tournament has at least one Hamiltonian path. Moreover, it is also possible to construct a tournament with only one Hamiltonian path.

Theorem. There is always a tournament on n vertices with at least

$$\frac{n!}{2^{n-1}}$$

Hamiltonian paths.

Definition. Given sets $S_1, S_2, \ldots, S_m \subseteq X$, we say that H is a *hitting set* if it intersects all the sets, i.e.,

$$H \cap S_i \neq \emptyset, \quad \forall i.$$

We will assume $|S_i| = k$.

Example.

- $H_1 = X_1, \ldots, X_3$.
- Any $H_2 \subseteq X_1, \ldots, X_3$.
- $|H_2| = n - k + 1$ by the pigeonhole principle.
- If H_1 is chosen randomly, then $H = X_{h_1}, \ldots, X_3$ where $h_i \in S_i$.

Theorem. Given sets $S_1, \ldots, S_r$, there exists a hitting set of size at most

$$\left\lceil \frac{n \log m}{k} \right\rceil.$$