Probability Notes

Nikhil R.

Contents

1 Experiments with Random Outcomes

1.1 The Probability Space

Definition (The Probability Space). A formal **probability space** is defined by the triplet $(\Omega, \mathcal{F}, \mathbb{P})$.

- **Sample Space (Ω):** The set of all possible outcomes.
- **Events ($\mathcal{F}$):** A collection of subsets of Ω (specifically, $\mathcal{F} \equiv \mathcal{P}(\Omega)$ in discrete cases).
  - Elementary Events: The individual elements of Ω.
- **Probability Distribution ($\mathbb{P}$):** A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ that assigns a probability to events and satisfies the following properties:
  - Range: For any event $A \in \mathcal{F}$, $\mathbb{P}(A) \in [0, 1]$.
  - Normalization: $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$.
  - Countable Additivity: If $\{A_i\}$ are disjoint sets (i.e., $A_i \cap A_j = \emptyset$), then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

1.2 Random Sampling and Counting

Definition (Uniform Sampling). When choosing objects uniformly at random from a finite set where each outcome has the same likelihood, the probability of an event A is given by $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$.

We may determine the size of the sample space $|\Omega|$ based on the sampling method.

1. **Sampling with Replacement (Ordered):**
   - We pick k items from a set $S = \{1, \ldots, n\}$.
   - The sample space is $\Omega \equiv S^k$.
   - Count: $|\Omega| = n^k$.
2. **Sampling without Replacement (Ordered):**
   - Order matters, and items are not replaced.
   - Count: $|\Omega| = \frac{n!}{(n-k)!}$ (Permutations).
3. **Sampling without Replacement (Unordered):**
   - Order does not matter, leading to fewer possibilities.
   - Count: $|\Omega| = \binom{n}{k} = \frac{n!}{k!(n-k)!}$ (Combinations).

Example (Urn Problem). Consider an urn with 10 balls: 5 red, 3 blue, 2 green. We pick 3 balls and want exactly 2 red and 1 blue.

- Total Outcomes: $\#\Omega = \binom{10}{3}$ (order does not matter).
- Event A: $\binom{5}{2}$ ways to choose red balls $\times \binom{3}{1}$ way to choose blue balls.
- Probability: $\mathbb{P}(A) = \frac{\binom{5}{2}\binom{3}{1}}{\binom{10}{3}} = \frac{1}{4}$.

1.3 Infinite Sample Spaces

For uncountably infinite spaces, formal theory requires more complex measure theory, but we consider countably infinite cases here.

Example (Geometric Distribution). Flipping a coin until it lands on heads for the first time after k flips.

- Probability: $\mathbb{P}(k) = \frac{1}{2^k}$.
- Normalization: $\mathbb{P}(\Omega) = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1$. Consequently, the probability of *never* rolling heads ($\mathbb{P}(\infty)$) is 0.

1.4 Consequences of Probability Rules

1. **Complements:** If $A \cap A^c = \emptyset$ and $A \cup A^c = \Omega$, then:
$$\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$$
2. **Monotonicity:** For events A, B, if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
   - *Proof:* We can write B as the disjoint union of A and the part of B not in A: $B = A \cup (A^c \cap B)$.
   $$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(A^c \cap B)$$
   Since $\mathbb{P}(A^c \cap B) \geq 0$, it follows that $\mathbb{P}(B) \geq \mathbb{P}(A)$.
3. **Inclusion-Exclusion Principle:**
   - 2 Events: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
     - *Proof:* Decompose the union into disjoint sets: $A \cup B = (A \cap B) \cup (A \setminus B) \cup (B \setminus A)$.
   - 3 Events: $\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(AB) - \mathbb{P}(BC) - \mathbb{P}(AC) + \mathbb{P}(ABC)$.
   - General Idea: Intersections are double-counted and subtracted; this uncounts deeper intersections which must be added back.

1.5 Random Variables

Definition. Given a sample space Ω, a **random variable** is a function $X : \Omega \to \mathbb{R}$.

- Inverse Images: For different values $k \in \mathbb{R}$, the sets $\{\omega \in \Omega : X(\omega) = k\}$ are pairwise disjoint, allowing us to sum probabilities.

Example. Rolling two dice where $\Omega = \{(i, j) : i, j \in \{1..6\}\}$. Let $X_1(i, j) = i$, $X_2(i, j) = j$, and $S = i + j$.

Definition (Probability Distribution). The probability distribution of X is the collection of probabilities $\mathbb{P}(X \in B)$ where $B \subseteq \mathbb{R}$.

Definition. A random variable is **discrete** if there exists a countably infinite set K such that $\sum_i \mathbb{P}(X = k_i) = 1$. For a discrete random variable, the **probability mass function** (PMF) is defined as $\mathbb{P}(k) \equiv \mathbb{P}(X = k)$ where k is in the range of X. Here, we consider individual outcomes $X = k \in \mathbb{R}$ rather than subsets.

Example. For the sum of two dice S, we calculate probabilities by summing outcomes:

$$\mathbb{P}(2 \leq S \leq 5) = \sum_{k \in [2,5]} \mathbb{P}(S = k) = \frac{1 + 2 + 3 + 4}{36} = \frac{10}{36}$$

2 Conditional Probability

2.1 Conditional Probability

Definition. Suppose we have events A and B where $\mathbb{P}(B) \neq 0$. We define the **conditional probability** of A given B as $\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)}$.

- Note: If we consider the sample space restricted to B (i.e., $\Omega = B$), this conditional probability satisfies the axioms of a probability measure.
- For discrete sample spaces, this is calculated as $\mathbb{P}(A|B) = \frac{|AB|}{|B|}$.

Example (Urn Problem). Consider an urn with 8 red balls and 4 blue balls. We draw 2 balls without replacement.

- Let $R_1 = \{\text{1st ball is red}\}$ and $R_2 = \{\text{2nd ball is red}\}$.
- The probability both are red is $\mathbb{P}(R_1 \cap R_2) = \mathbb{P}(R_1)\mathbb{P}(R_2|R_1) = (\frac{8}{12})(\frac{7}{11}) = \frac{14}{33}$.
- Now, suppose we draw 4 balls total (R_1, R_2, W_3, W_4 where W is blue).
- $\mathbb{P}(R_1R_2W_3W_4) = \mathbb{P}(R_1)\mathbb{P}(R_2|R_1)\mathbb{P}(W_3|R_1R_2)\mathbb{P}(W_4|R_1R_2W_3) = (\frac{8}{12})(\frac{7}{11})(\frac{4}{10})(\frac{3}{9}) = \frac{28}{495}$.

Theorem (Chain Rule). More generally, we can calculate the probability of a sequence of events as:

$$\mathbb{P}(X_1 \dots X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_1X_2)\dots\mathbb{P}(X_n|X_1\dots X_{n-1})$$

2.2 Law of Total Probability and Bayes' Rule

Theorem (Law of Total Probability). Suppose the sets $B_1, \dots, B_n$ form a partition of Ω (meaning their union is Ω and they are pairwise disjoint) where $\mathbb{P}(B_i) > 0$. Then for any event $A \subseteq \Omega$:

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i) = \sum_i \mathbb{P}(A|B_i)\mathbb{P}(B_i)$$

Example (Biased Coins). Suppose we have 10 coins: 9 are fair ($\mathbb{P}(T) = 0.5$) and 1 is biased ($\mathbb{P}(T) = 0.9$). If we pick a coin at random and flip it:

- $\mathbb{P}(T) = \mathbb{P}(T|F)\mathbb{P}(F) + \mathbb{P}(T|M)\mathbb{P}(M)$ (where F is fair, M is "moderate" or biased).
- Calculation: $(\frac{1}{2})(0.9) + (\frac{9}{10})(0.1) = 0.54$.

Theorem (Bayes' Rule). Suppose $B_1, \dots, B_n$ partition Ω and $\mathbb{P}(A), \mathbb{P}(B_i) > 0$. Then:

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{i=1}^n \mathbb{P}(A|B_i)\mathbb{P}(B_i)}$$

- *Remark:* The first step is completely symmetric, and the second step uses the Law of Total Probability.

Example (Coins continued). Using the previous example, suppose we get tails (T). What is the probability we picked the fair/moderate coin (F)?

- $\mathbb{P}(F|T) = \frac{\mathbb{P}(T|F)\mathbb{P}(F)}{\mathbb{P}(T)} = \frac{(0.5)(0.9)}{0.54} \approx 0.833$.

Example (Disease Testing). 0.5% of people carry a disease. A test detects the disease 96% of the time but has a 2% false positive rate.

- Given: $\mathbb{P}(+|D) = 0.96$, $\mathbb{P}(+|D^c) = 0.02$, $\mathbb{P}(D) = 0.005$, $\mathbb{P}(D^c) = 0.995$.
- We want $\mathbb{P}(D|+)$ (Probability of disease given positive test).
- Calculation: $\mathbb{P}(D|+) = \frac{\mathbb{P}(+|D)\mathbb{P}(D)}{\mathbb{P}(+|D)\mathbb{P}(D)+\mathbb{P}(+|D^c)\mathbb{P}(D^c)} = \frac{(0.96)(0.005)}{(0.96)(0.005)+(0.02)(0.995)} \approx 0.194$.
- *Observation:* This result is very sensitive to false positives because $\mathbb{P}(D)$ is incredibly small. If $\mathbb{P}(D) = 0.5$, the result would be much higher (≈ 0.98).

2.3 Independence

Definition. Events A and B are **independent** ($A \perp B$) if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

- *Proof of Consistency:* $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$. If $A \perp B$, then $\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A|B)\mathbb{P}(B)$, which implies $\mathbb{P}(A|B) = \mathbb{P}(A)$. This matches the intuition that knowing B gives no information about A.

Example (Coin Flips). Flipping a fair coin 3 times.

- $A = \{\text{exactly 1 T in first 2 flips}\} = \{(H,T,T),(H,T,H),(T,H,T),(T,H,H)\}$.
- $B = \{\text{exactly 1 T in last 2 flips}\}$.
- $C = \{\text{exactly 1 T total}\}$.
- Sample Space size $\#\Omega = 2^3 = 8$. $\mathbb{P}(A) = 1/2$, $\mathbb{P}(B) = 1/2$, $\mathbb{P}(C) = 3/8$.
- Intersection: $A \cap B = \{(T,H,T),(H,T,H)\}$, so $\mathbb{P}(A \cap B) = 2/8 = 1/4$.
- Check: $\mathbb{P}(A)\mathbb{P}(B) = (1/2)(1/2) = 1/4$. Thus, A and B are **independent**.
- Check C: $A \cap C$ and $B \cap C$ calculations would show they are not independent.

Example (Sampling). Urn with 4 red, 7 green balls. Choose 2 balls.

- **With Replacement:** $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Events are independent.
- **Without Replacement:** $\mathbb{P}(A) = 4/11$. $\mathbb{P}(B)$ requires total probability. $\mathbb{P}(A \cap B) = (4/11)(7/10)$. Since the probabilities shift, the events are **not independent**.

Theorem (Independence of Complements). If $A \perp B$, then $A \perp B^c$, $A^c \perp B$, and $A^c \perp B^c$.

- *Proof:* Suppose $A \perp B$. We know $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(A^c \cap B)$. $\mathbb{P}(A^c \cap B) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)(1 - \mathbb{P}(A)) = \mathbb{P}(B)\mathbb{P}(A^c)$. Thus $A^c \perp B$. The proof works similarly for the other cases.

2.4 Mutual vs Pairwise Independence

Definition. Events $A_1, \ldots, A_n$ are **mutually independent** if for all subsets of indices $K \subseteq \{1, \ldots, n\}$, the probability of the intersection equals the product of the probabilities:

$$\mathbb{P}\left(\bigcap_{k \in K} A_k\right) = \prod_{k \in K} \mathbb{P}(A_k)$$

Definition. Events are **pairwise independent** if $A_i \perp A_j$ for all $i \neq j$.

- *Note:* Mutual independence is a stronger condition that implies pairwise independence, but pairwise does not imply mutual.

Example (Intervals). Choose a real number from $[0, 1]$.

- Let $A = [0, 1/2]$, $B = [1/4, 3/4]$, $C = [0, 1/4] \cup [1/2, 3/4]$.
- Intersections: $A \cap B = [1/4, 1/2]$ (length $1/4$). $B \cap C$ (length $1/4$). $A \cap C$ (length $1/4$).
- Probabilities: $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = 1/2$.
- Pairwise: $\mathbb{P}(A \cap B) = 1/4 = \mathbb{P}(A)\mathbb{P}(B)$. Same for others. They are pairwise independent.
- Mutual: $A \cap B \cap C = \emptyset$ (or single points of measure 0). Thus $\mathbb{P}(A \cap B \cap C) = 0$.
- Check: $\mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C) = 1/8 \neq 0$. They are **not** mutually independent.

Definition. Random variables $X_1, \ldots, X_n$ are **independent** if for all subsets of subsets $B_1, \ldots, B_n \subseteq \mathbb{R}$:

$$\mathbb{P}(\{X_i \in B_i : i = 1 \ldots n\}) = \prod_{k=1}^{n} \mathbb{P}(X_k \in B_k)$$

where $\{X_k \in B_k\} = \{\omega \in \Omega : X_k(\omega) \in B_k\}$.

2.5 Independent Trials and Distributions

Definition. A sequence of trials is **independent** if the outcome of one does not affect the others. If a trial is repeated n times with outcomes success (1) or failure (0), the sample space is the binary field $\Omega = \{0, 1\}^n$. The probability of a specific sequence ω with k successes is $\mathbb{P}(\omega) = p^k(1-p)^{n-k}$.

Definition. A random variable X has a **Bernoulli distribution**, denoted $X \sim \text{Bern}(p)$, if it takes values in $\{0, 1\}$ with success probability $\mathbb{P}(\text{success}) = p$ and failure probability $\mathbb{P}(\text{failure}) = 1 - p$.

Definition. A random variable X has a **Binomial distribution**, denoted $X \sim \text{Bin}(n, p)$, if it counts the number of successes in n independent Bernoulli trials with parameter p.

- Let $X_1, \ldots, X_n$ be independent Bern(p) variables; then $S_n \equiv \sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$.
- The probability mass function is:

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k \in \{0, \ldots, n\}$$

- *Normalization Proof:* $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1^n = 1$ (by Binomial Theorem).

Example (Dice Rolls). Rolling a fair die 5 times. Probability of getting an odd number of 6s.

- Success (6) has $p = 1/6$.
- We sum for $k \in \{1, 3, 5\}$: $\mathbb{P}(\text{odd } \#6) = \sum_{k \in \{1,3,5\}} \binom{5}{k} (1/6)^k (5/6)^{5-k}$.

Definition. A random variable X has a **Geometric distribution**, denoted $X \sim \text{Geom}(p)$, if it counts the number of trials needed to get the first success in a sequence of independent Bernoulli trials.

- Range: $X \in \mathbb{N}$ (i.e., $1, 2, \ldots$).
- PMF: $\mathbb{P}(X = k) = (1-p)^{k-1} p$.
- *Interpretation:* $k - 1$ failures followed by 1 success.

Definition. A random variable X has a **Hypergeometric distribution**, denoted $X \sim \text{Hypergeom}(N, N_A, n)$, if it counts the number of type A items in a sample of size n drawn without replacement from a population of size N.

- Population: N_A (Type A) + N_B (Type B) = N.
- Range: $k \in \{\max(0, n - N_B), \ldots, \min(n, N_A)\}$.
- PMF: $\mathbb{P}(X = k) = \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}}$

3 Random Variables

3.1 Densities and Cumulative Distribution Functions

Definition. The **cumulative distribution function** (CDF) for a random variable X is defined as $F(a) \equiv \mathbb{P}(X \leq a)$.

Definition. A random variable is **discrete** if CDF is defined as $F(s) = \mathbb{P}(X \leq s) = \sum_{k \leq s} \mathbb{P}(X = k)$. The graph of the CDF for a discrete RV is a piecewise, step-constant function. At possible outcomes of the RV, the height of the jump (size of discontinuity) corresponds to the probability of that outcome.

Definition. The RV X has a **continuous distribution** if there exists a function $f_x : \mathbb{R} \to [0, \infty)$ such that for all $a \in \mathbb{R}$ $\mathbb{P}(X \leq a) = \int_{-\infty}^{a} f_x(x)dx$. Here, f_x is the called the **probability density function** (PDF).

Proposition. For a continuous random variable:

- If $B \subseteq \mathbb{R}$, then $\mathbb{P}(X \in B) = \int_B f_x(x)dx$.
- $\mathbb{P}(X = a) = \int_a^a f_x(x)dx = 0$.
- $\mathbb{P}(X < a) + \mathbb{P}(X = a) = \mathbb{P}(X \leq a)$.
- Normalization: $\int_{-\infty}^{\infty} f_x(x)dx = \mathbb{P}(X \in \mathbb{R}) = 1$.
- Intervals: $\mathbb{P}(a \leq X \leq b) = \int_{-\infty}^{b} f_x dx - \int_{-\infty}^{a} f_x dx = \int_a^b f_x dx$.

Proposition. For a general distribution (continuous or discrete):

- Monotonicity: If $s < t$, then $F(s) \leq F(t)$.
- Right Continuity: $F(t) = \lim_{s \to t^+} F(s)$.
- Endpoints: $\lim_{s \to -\infty} F(s) = 0$ and $\lim_{s \to \infty} F(s) = 1$.
- Left limits: $\mathbb{P}(X < a) = \lim_{s \to a^-} F(s)$.
- Interval probability: $\mathbb{P}(a < X \leq b) = F(b) - F(a)$.
- CDF to PDF relation: $F_x'(a) = \frac{d}{da}(\int_{-\infty}^{a} f_x(x)dx) = f_x(a)$.

Definition. Let $[a, b] \subseteq \mathbb{R}$ be compact. The RV X has the **uniform distribution** on $[a, b]$ ($X \sim \text{Unif}([a, b])$) if the PDF is $f_x(x) = \frac{1}{b-a}$ for $x \in [a, b]$, and 0 otherwise.

3.2 Expectation Values

Definition. If X is a discrete RV with values k, then $\mu = \mathbb{E}(X) = \sum_k k\mathbb{P}(X = k)$. If X is a continuous RV with PDF $f(x)$, then $\mu = \mathbb{E}(X) = \int_{\mathbb{R}} xf(x)dx$.

Proposition.

- $X \sim \text{Ber}(p)$, then $\mu = p$
  - *Proof:* $\mu = \mathbb{E}(X) = 1 \cdot p + 0 \cdot (1-p) = p$
- $X \sim \text{Bin}(n, p)$, then $\mu = np$
  - *Proof:*

$$\begin{aligned}\mu &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!}p^k(1-p)^{n-k} \\ &= np\sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!}p^{k-1}(1-p)^{n-k} \\ &= np\sum_{j=0}^{n-1} \frac{(n-1)!}{j!((n-1)-j)!}p^j(1-p)^{(n-1)-j} = np(1) = np\end{aligned}$$

- $X \sim \text{Geom}(p)$, then $\mu = \frac{1}{p}$
  - *Proof:*

$$\begin{aligned}\mu = \mathbb{E}(X) &= \sum_{k=1}^{\infty} k(1-p)^{k-1}p = -p\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^k \\ &= -p\frac{d}{dp}\left(\frac{1}{1-(1-p)} - 1\right) = -p\frac{d}{dp}\left(\frac{1}{p} - 1\right) = -p\left(-\frac{1}{p^2}\right) = \frac{1}{p}\end{aligned}$$

- $X \sim \text{Unif}([a, b])$, then $\mu = \frac{a+b}{2}$
  - *Proof:* $\mu = \int_{\mathbb{R}} xf(x)dx = \int_a^b x\frac{1}{b-a}dx = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{a+b}{2}$

Example. Consider a dartboard with density $f(r) = 2r/R^2$ for $r \in [0, R]$. Then, $\mu = \int_0^R r\frac{2r}{R^2}dr = \frac{2}{R^2}\int_0^R r^2dr = \frac{2}{3}R$

Example (St. Petersburg Paradox). Flipping a coin where tails doubles the prize (2^n). $\mathbb{P}(X = 2^n) = \frac{1}{2^n}$. Then $\mathbb{E}(X) = \sum_{n=1}^{\infty} 2^n\frac{1}{2^n} = \sum 1 = \infty$

Example. Flipping a coin. If the first head is on an odd flip, I lose 2^n; on even, I make 2^n. $\mu = \sum_n^{\infty}\left(-2^{2n}\frac{1}{2^{2n}} + 2^{2n+1}\frac{1}{2^{2n+1}}\right)$. This series does not converge.

3.3 Functions of Random Variables

Proposition. Suppose $X : \Omega \to \mathbb{R}$ and $g : \text{Range}(X) \to \mathbb{R}$. Then,

- Discrete Case: $\mathbb{E}(g(X)) = \sum_k g(k)\mathbb{P}(X = k)$

- *Proof:* $g(X(\omega)) = y$ is a partition for Ω so $\mathbb{E}(g(X)) = \sum_y y\mathbb{P}(g(X) = y) = \sum_y y \sum_{k:g(k)=y} \mathbb{P}(X = k) = \sum_k g(k)\mathbb{P}(X = k)$

- Continuous Case: $\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx$

Example. Let $g(X) = X^n$. For $X \sim \text{Unif}[0, c]$ we get $\mathbb{E}(X^n) = \int_0^c u^n \frac{1}{c} du = \frac{1}{c}\frac{c^{n+1}}{n+1} = \frac{c^n}{n+1}$

Example (Car Insurance). Let Bill $Y \sim \text{Unif}[100, 1500]$. You pay $g(Y) = \min(Y, 500)$. Then $\mu = \int_{100}^{500} Y \frac{1}{1400} dY + \int_{500}^{1500} 500 \frac{1}{1400} dY \approx 443$

3.4 Median and Quartiles

Definition. The **median** of an RV, $m \in \mathbb{R}$, is such that $\mathbb{P}(X \geq m) \geq 1/2$ and $\mathbb{P}(X \leq m) \geq 1/2$. This represents the midpoint while weeding out outliers.

Example. For $S = \{-100, 1, 2, ..., 9\}$, median is 4.5, while mean is near -4.5.

Definition. The **p-th quartile**, $q \in \mathbb{R}$, is such that $\mathbb{P}(X \leq q) \geq p$ and $\mathbb{P}(X \geq q) \geq 1 - p$.

- 1st quartile ($p = 0.25$) and 3rd quartile ($p = 0.75$).

3.5 Variance

Definition. The **variance** of an RV X is a measure of how much X fluctuates around its mean

$$\sigma^2 = \text{Var}(X) = \mathbb{E}((X - \mu)^2)$$

- Discrete: $\sigma^2 = \sum_k (k - \mu)^2 \mathbb{P}(X = k)$

- Continuous: $\sigma^2 = \int_{\mathbb{R}} (x - \mu)^2 f(x)dx$

Proposition (Linearity). $\mathbb{E}(ax + b) = a\mathbb{E}(x) + b$ and $\sigma^2(ax + b) = a^2\sigma^2(x)$

- *Proof:* Linearity of expectation comes from the linearity of the integral and sums. For variance, $\mathbb{E}((ax + b - \mathbb{E}(ax + b))^2) = \mathbb{E}((ax - a\mathbb{E}(x))^2) = a^2\mathbb{E}((x - \mathbb{E}(x))^2) = a^2\sigma^2(x)$

Proposition. $\sigma^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$

- *Proof:* $\sigma^2 = \mathbb{E}(X^2 - 2\mu X + \mu^2) = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$

Proposition.

- $X \sim \text{Ber}(p)$, then $\sigma^2 = p(1 - p)$

  - *Proof:* $\sigma^2 = (1 - p)^2 p + p^2(1 - p) = (1 - p)(p - p^2 + p^2) = p(1 - p)$

- $X \sim \text{Bin}(p)$, then $\sigma^2 = np(1 - p)$

  - *Proof:* $\mathbb{E}(X^2) = \sum_{k=1}^n k^2 \binom{n}{k} p^k (1 - p)^{n-k}$. Writing $k^2 = k(k - 1) + k$, we get $\mathbb{E}(X^2) = n(n - 1)p^2 + np + \mathbb{E}(X)$. Thus $\sigma^2 = (n(n - 1)p^2 + np) - (np)^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1 - p)$

- $X \sim \text{Unif}([a,b])$, then $\sigma^2 = \frac{(b-a)^2}{12}$
  - *Proof:* $\mathbb{E}(X^2) = \frac{1}{b-a}\int_a^b x^2 dx = \frac{1}{3}(b^2+ab+a^2)$ so $\sigma^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$
- $X \sim \text{Geom}(p)$, then $\sigma^2 = \frac{1-p}{p^2}$
  - *Proof:*

$$\begin{aligned}\sigma^2 &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 = \left[\sum_{k=1}^{\infty} k(k-1)(1-p)^{k-1}p\right] + \frac{1}{p} - \frac{1}{p^2} \\ &= p(1-p)\left[\frac{d^2}{dq^2}\sum_{k=0}^{\infty} q^k\right]_{q=1-p} + \frac{p}{p^2} - \frac{1}{p^2} = p(1-p)\left[\frac{2}{(1-(1-p))^3}\right] + \frac{p-1}{p^2} \\ &= p(1-p)\left(\frac{2}{p^3}\right) - \frac{1-p}{p^2} = \frac{2(1-p)}{p^2} - \frac{1-p}{p^2} = \frac{1-p}{p^2}\end{aligned}$$

Proposition. $\sigma^2(X) = 0$ if and only if there exists k such that $\mathbb{P}(X = k) = 1$.

- *Proof:* ($\Leftarrow$) If $\mathbb{P}(X = k) = 1$, then $\mathbb{E}(X) = k$ and $\mathbb{E}((k-k)^2) = 0$. ($\Rightarrow$) If $\sum(k-\mu)^2\mathbb{P}(X = k) = 0$, since terms are non-negative, $k-\mu$ must be 0 for all k with non-zero probability. Thus $k = \mu$ with probability 1.

3.6 Gaussian Distribution

Definition. An RV Z has a **standard normal distribution** ($Z \sim N(0,1)$) if the PDF is $f_z = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$

- Symmetric bell curve. Concave down near $z = 0$, and has inflection points at $z = \pm 1$
- If $I = \int_{-\infty}^{\infty} e^{-x^2}dx$, $I^2 = \iint_{\mathbb{R}^2} e^{-(x^2+y^2)}dxdy = \int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2} r dr = 2\pi\left[-\frac{1}{2}e^{-r^2}\right]_0^{\infty} = \pi \Rightarrow I = \sqrt{\pi}$, thus, for $N(0,1)$, the factor $\frac{1}{\sqrt{2\pi}}$ ensures the integral is 1.

Proposition. For $Z \sim N(0,1)$, define $X = |\sigma|Z + \mu$. Then $X \sim N(\mu, \sigma^2)$

- *Proof:* $\mathbb{E}(X) = \mu + |\sigma|\mathbb{E}(Z) = \mu$ and $\text{Var}(X) = \sigma^2 Var(Z) = \sigma^2$
- $F_x(a) = \mathbb{P}(\mu + \sigma Z \le a) = \mathbb{P}(Z \le \frac{a-\mu}{|\sigma|}) = \Phi(\frac{a-\mu}{|\sigma|})$
- $f_x(a) = \frac{d}{da}\Phi\left(\frac{a-\mu}{|\sigma|}\right) = f_z\left(\frac{a-\mu}{|\sigma|}\right)\frac{1}{|\sigma|} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(a-\mu)^2}{2\sigma^2}\right)$

4 Approximations of Binomial Distribution

4.1 Central Limit Theorem

Proposition. Suppose $n \gg 1$ and for $p \in (0,1)$, let $S_n \sim \text{Bin}(n,p)$. We previously computed that $\mathbb{E}(S_n) = np$ and $\sigma^2(S_n) = np(1-p)$. We empirically note that as n becomes large, the shape of the pmf of $\text{Bin}(n,p)$ (which is discrete) approaches the Gaussian distribution. In particular, suppose $-\infty \le a \le b \le +\infty$. Then:

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Thus $S_n \sim \text{Bin}(n,p)$ approaches $X \sim N(np, np(1-p))$ for $n \to \infty$. (In the limit, we can equate probabilities of different values of S_n, but not the pmf to Gaussian pdf since one is discrete and the other is continuous.)

- *Proof Sketch:*
  - We start with Stirling's formula, $n! \approx \sqrt{2\pi n}(n/e)^n$. We approximate the probability mass function $\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$. Let $q = 1-p$. Then,

$$\begin{aligned}
\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} &= \frac{n!}{k!(n-k)!} p^k q^{n-k} \\
&\approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}} p^k q^{n-k} \\
&= \frac{\sqrt{2\pi n}}{\sqrt{2\pi k}\sqrt{2\pi(n-k)}} \cdot \frac{(n/e)^n}{(k/e)^k((n-k)/e)^{n-k}} \cdot p^k q^{n-k} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \frac{n^n}{k^k (n-k)^{n-k}} \cdot p^k q^{n-k} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{n^k p^k}{k^k}\right)\left(\frac{n^{n-k} q^{n-k}}{(n-k)^{n-k}}\right) \quad \text{by } n^n = n^k n^{n-k} \\
&= \sqrt{\frac{n}{2\pi k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k}
\end{aligned}$$

  - Let the deviation from the mean be $\delta = k - np$, where we assume $\frac{\delta}{np} \ll 1$. Then $k = np + \delta$ and $n - k = nq - \delta$. Let $L = \ln[(\frac{np}{k})^k (\frac{nq}{n-k})^{n-k}]$. Substituting δ, we get $L = -\left(k\ln(1+\frac{\delta}{np}) + (n-k)\ln(1-\frac{\delta}{nq})\right)$. Using the Taylor expansion $\ln(1+x) \approx x - x^2/2$ for small x, we expand terms to the second order. The first-order terms cancel out, leaving $L \approx -\frac{\delta^2}{2np} - \frac{\delta^2}{2nq} = -\frac{\delta^2}{2npq}$.
  - For the coefficient, using $k = np + \delta$, we get $\sqrt{\frac{n}{2\pi k(n-k)}} = \sqrt{\frac{n}{2\pi(np+\delta)(nq-\delta)}} \approx \frac{1}{\sqrt{2\pi npq}}$.
  - Exponentiating L back gives $\mathbb{P}(S_n = k) \approx \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(k-np)^2}{2npq}}$, which we note is the pdf of $X \sim N(np, npq)$

Example. Rolling 1, 2 moves up 2. Rolling 3, 4, 5, 6 moves up 3. Calculate $\mathbb{P}$(moves $>$ 315 after 120 rolls).

- Let $X_n \equiv$ Moves up after n rolls, $A_n \equiv \#$ of 1, 2 rolled in n trials, $B_n \equiv \#$ of 3, 4, 5, 6 rolled. Then $A_n \sim \text{Bin}(n, 1/3)$ and $B_n = n - A_n$
- $X_n = 2A_n + 3(n - A_n) = 3n - A_n$
- $\mathbb{E}(X_n) = 3n - \mathbb{E}(A_n) = 3n - np = (3-p)n = \frac{8}{3} \cdot 120 = 320$
- $\sigma^2(X_n) = \sigma^2(A_n) = np(1-p) = 120 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{80}{3}$
- By the CLT, $\mathbb{P}(X_n > 315) = \mathbb{P}\left(Z > \frac{315-320}{\sqrt{80/3}}\right) = 1 - \Phi\left(\frac{-5}{\sqrt{80/3}}\right) = \Phi\left(\frac{5\sqrt{3}}{\sqrt{80}}\right) \approx 0.834$

4.2 Law of Large Numbers

Proposition. (We will later prove this for more general iid RV's.) Suppose there are $X_1, X_2, ... \sim \text{Ber}(p)$ be independent trials, each with probability p. Let $S_n = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$. Then $p \in (0, 1)$. Then,

$$\lim_{n\to\infty} \mathbb{P}\left(\epsilon > \left|\frac{S_n - np}{n}\right|\right) = 1$$

This is a "weak" law. The stronger formulation would be demanding that $\mathbb{P}(\lim S_n/n = p) = 1$ (that the probability of the random variable S_n/n itself converges to p, not just deviations become unlikely).

- *Proof:* $\mathbb{P}\left(\epsilon > \left|\frac{S_n}{n} - p\right|\right) = \mathbb{P}(-n\epsilon < S_n - np < n\epsilon) = \mathbb{P}\left(\frac{-\epsilon n}{\sigma} < \frac{S_n - np}{\sigma} < \frac{\epsilon n}{\sigma}\right)$. But by CLT, this is $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1$ in the limit.

Example. Flipping a coin. Let $S_n = \#$ of tails in n flips. $S_n \sim \text{Bin}(n, 1/2)$. Then, $\mathbb{P}(S_n/n \geq 0.51) = \mathbb{P}(S_n/n - 0.5 \geq 0.01)$. We can bound this by considering the absolute value, for which there are two cases instead of just one case. Thus, our desired is $\leq \mathbb{P}\left(\left|\frac{S_n}{n} - 0.5\right| \geq 0.01\right) = 1 - \mathbb{P}\left(\left|\frac{S_n}{n} - 0.5\right| < 0.01\right) = 1 - 1 = 0$ by LLN.

4.3 Confidence Intervals for Binomial Distributions

Proposition. Our goal is to estimate p by the Law of Large Numbers. Suppose we take n trials and get S_n successes for a Bernoulli RV. Then let, $p^\dagger = S_n/n$. How far off is this from

actual p? Suppose $\epsilon > 0$.

$$\begin{aligned}
\mathbb{P}(|p^\dagger - p| < \epsilon) &= \mathbb{P}(-n\epsilon < S_n - np < n\epsilon) \\
&= \mathbb{P}\left(\frac{-n\epsilon}{\sqrt{np(1-p)}} < \frac{S_n - np}{\sqrt{np(1-p)}} < \frac{n\epsilon}{\sqrt{np(1-p)}}\right) \\
&= \Phi\left(\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{-\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right) \\
&= 2\Phi\left(\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right) - 1 \quad \text{by symmetry} \\
&\geq 2\Phi(2\sqrt{n}\epsilon) - 1 \quad \text{since } \max(p(1-p))|_{[0,1]} = \frac{1}{4}
\end{aligned}$$

4.4 Poisson Distribution

Definition. Let $\lambda > 0$. The RV X has **Poisson Distribution** with parameter lambda ($X \sim \text{Poisson}(\lambda)$) if X takes values in $0, 1, 2, ...$ and has pmf:

$$\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \text{ for } k \in \{0, 1, ...\}$$

- Normalization: $\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}(\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}) = e^{-\lambda}e^{\lambda} = 1$
- Expectation: $\mathbb{E}(X) = \lambda$
  - *Proof:* $\mathbb{E}(X) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} = \mu = \lambda$
- Variance: $\text{Var}(X) = \lambda$
  - *Proof:* Noting that $\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + \mathbb{E}(X)$, $\sigma^2 + \mu^2 - \mu = \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2 e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2$ meaning $\sigma^2 = \lambda$

Proposition. The Poisson Distribution is a good model for occurrences that are independent from each other, where probability of success is constrained by $np = \lambda$. In particular, suppose $S_n \sim \text{Bin}(n, p)$ and $np = \lambda$ for $\lambda > 0$. Then, $\lim_{n\to\infty}\mathbb{P}(S_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}$

- *Proof:*

$$\begin{aligned}
\mathbb{P}(S_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n(n-1)(n-2)...(n-(k-1))}{k!}\frac{\lambda^k}{n^k}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k} \\
&= \left(1 - \frac{\lambda}{n}\right)^n\frac{\lambda^k}{k!}\frac{1(1-1/n)(1-2/n)...(1-(k-1)/n)}{(1-\lambda/n)^k} \\
&\to e^{-\lambda}\frac{\lambda^k}{k!}\frac{(1\cdot 1\cdot ...\cdot 1)}{1} = e^{-\lambda}\frac{\lambda^k}{k!}
\end{aligned}$$

4.5 Exponential Distribution

Definition. Let $\lambda \in (0, \infty)$. An RV X has the **exponential distribution**, denoted by $X \sim Exp(\lambda)$ if the pdf is $f_x(y) = \begin{cases} 0, & y < 0 \\ \lambda e^{-\lambda y}, & y \geq 0 \end{cases}$

- CDF: $F_x(a) = \int_0^a \lambda e^{-\lambda y} dy = -[e^{-\lambda y}]_0^a = 1 - e^{-\lambda a}$ for $a \geq 0$
- Expectation: $\mu = \frac{1}{\lambda}$
  - *Proof:* $\mathbb{E}(X) = \int_0^\infty y\lambda e^{-\lambda y} dy = \left[\frac{-(1+\lambda y)}{\lambda} e^{-\lambda y}\right]_0^\infty = \frac{1}{\lambda}$
- Variance: $\sigma^2 = \frac{1}{\lambda^2}$
  - *Proof:* $\mathbb{E}(X^2) = \int_0^\infty y^2 \lambda e^{-\lambda y} dy = \left[\frac{-((\lambda y)^2 + 2\lambda y + 2)e^{-\lambda y}}{\lambda^2}\right]_0^\infty = \frac{2}{\lambda^2}$ so $\sigma^2 = \mathbb{E}(X^2) - \mu^2 = \frac{1}{\lambda^2}$

Proposition (Memoryless Property). Things behave as if they were brand new. Suppose $X \sim \text{Exp}(\lambda)$, $s, t > 0$. Then, $\mathbb{P}(X > t) = 1 - F_x(t) = e^{-\lambda t}$ and $\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s)$

- *Proof:* $\frac{\mathbb{P}(X > t+s \cap X > t)}{\mathbb{P}(X > t)} = \frac{\mathbb{P}(X > t+s)}{\mathbb{P}(X > t)} = \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s}$

Remark. The geometric distribution also satisfies the memoryless property.

Example. Calculate time for a protein to degrade given by $T \sim \text{Exp}(2n)$. Then, $\mathbb{P}(T > 1/100) < 1/10$, so $1/10 > \int_{1/100}^\infty f_T(y) dy = -e^{-2ny}]_{1/100}^\infty = e^{-2n/100}$. Thus, $\frac{-2n}{100} < \ln(1/10) \Rightarrow n > 50 \ln(10) \approx 115.1,$ so $n = 116$

Example. Suppose the car has already done 8000 miles. What is the probability that you are able to do another 5000 miles without the battery dying? Assume $X \sim \text{Exp}(\lambda)$, $\mathbb{E}(X) = 10^4$ miles. Since for the exponential distribution $\frac{1}{\lambda} = \mathbb{E}(X)$, we get $\mathbb{P}(\text{success}) = 1 - \mathbb{P}(\text{fail}) = 1 - \int_0^{5000} \lambda e^{-\lambda y} dy = e^{-\lambda(5000)} = e^{-10^{-4}(5000)} = e^{-1/2} \approx 0.6065$

5 Transforms and Transformations

5.1 Moment Generating Function

Definition. The **Moment Generating Function** (mgf) of a random variable X is given by $M_x(t) = \mathbb{E}(e^{tX})$ for $t \in \mathbb{R}$.

Proposition. Suppose $\exists\ t > 0$ such that $M_x(t)$ is bounded for $t \in (-\epsilon, +\epsilon)$. Then, we get the n^{th} moment of X by differentiating:

$$\langle X^n \rangle = \left[\mathbb{E}\left(\frac{d^n}{dt^n}(e^{tX}) \right) \right]_{t=0} = \left[\frac{d^n}{dt^n}(M_x(t)) \right]_{t=0}$$

Proposition. $(X \sim \text{Exp}(\lambda))$

$$\begin{aligned} M_X(t) &= \int_0^\infty e^{ty} \lambda\ e^{-\lambda y} dy = \lambda \int_0^\infty e^{(t-\lambda)y} dy \\ &= \left[\frac{\lambda}{t-\lambda} e^{(t-\lambda)y} \right]_0^\infty = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda \end{aligned}$$

Proposition. $(Z \sim N(0,1))$

$$\begin{aligned} M_Z(t) = \mathbb{E}(e^{tZ}) &= \int_{-\infty}^\infty e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-z^2/2 + tz} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{(z-t)^2}{2} + \frac{t^2}{2}} dz = e^{t^2/2} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{(z-t)^2}{2}} dz \right) \\ &= e^{t^2/2} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-u^2/2} du \right) \Bigg|_{u=z-t} = e^{t^2/2} \end{aligned}$$

Proposition. For $Z \sim N(0,1)$, the odd moments are 0 and for $n = 2k$, $\mathbb{E}(Z^{2k}) = \frac{(2k)!}{2^k k!}$

- *Proof:* We expand $\sum_{n=0}^\infty \frac{\mathbb{E}(Z^n)}{n!} t^n = \sum_{k=0}^\infty \frac{t^{2k}}{2^k k!}$ and equate coefficients.

5.2 Equating Distributions

Definition. Two RVs X and Y are **equal in distribution** (denoted $X =_d Y$) if $(\forall\ B \subseteq \mathbb{R})$ $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$. In particular, we note that this allows for the domain of X and the domain of Y to be different sample spaces.

Proposition (Criteria for Equality).

- Discrete RVs: $X =_d Y$ if and only if $\text{pmf}(X) = \text{pmf}(Y)$.
- Continuous RVs: $X =_d Y$ if and only if $\text{pdf}(X) = \text{pdf}(Y)$.
- Generally: $X =_d Y$ if and only if $\text{cdf}(X) = \text{cdf}(Y)$ or if $\forall\ g : \mathbb{R} \to \mathbb{R}$,
- Generally, $X =_d Y$ if and only if $(\forall\ g : \mathbb{R} \to \mathbb{R})$ $\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$.

- MGF: If for $t \in [-\delta, \delta]$, $M_x(t) = M_y(t) < \infty$, then $X =_d Y$.

Example. Suppose $M_x(t) = \frac{1}{5}e^{-17t} + \frac{1}{4} + \frac{11}{20}e^{2t} = \mathbb{E}(e^{tX})$. Then X is equal in distribution to a discrete RV which takes on value -17, 0, and 2 with respective probabilities $\frac{1}{5}, \frac{1}{4}, \frac{11}{20}$.

5.3 Distributions Functions of an RV

Proposition. Suppose X is a discrete random variable defined on Ω and that $Y = g(X)$ for some function $g : \mathbb{R} \to \mathbb{R}$. Then Y is also a discrete random variable with

$$\text{pmf}_Y(l) = \mathbb{P}(g(X) = l) = \sum_{k:g(k)=l} \text{pmf}_X(k)$$

- *Proof:* Let $A_l = \{k \in \text{Range}(X) : g(k) = l\}$ be the pre-image of l under g. The event $\{Y = l\}$ is equivalent to the event $\{X \in A_l\}$. Since the outcomes $X = k$ are disjoint events for distinct k, we can apply the countable additivity of the probability measure:

$$\mathbb{P}(Y = l) = \mathbb{P}\left(\bigcup_{k \in A_l} \{X = k\}\right) = \sum_{k \in A_l} \mathbb{P}(X = k)$$

  In particular, if g is injective (one-to-one), then A_l contains at most one element $x_0 = g^{-1}(l)$, and the sum collapses to a single term $\text{pmf}_X(g^{-1}(l))$.

Proposition. Let X be a continuous RV with CDF F_X. Let $Y = g(X)$. It is not immediately known whether Y is continuous or discrete. The CDF of Y is generally determined by finding the set $A_y = \{x : g(x) \leq y\}$ and computing $F_Y(y) = \mathbb{P}(X \in A_y)$. However, we may consider the special case where g^{-1} exists, meaning g is strictly monotone.

- Increasing: $\{g(X) \leq y\} \iff \{X \leq g^{-1}(y)\}$. Thus $F_Y(y) = F_X(g^{-1}(y))$.
  - *Proof:* Since g is increasing, $u \leq v \iff g(u) \leq g(v)$. Applying g^{-1} preserves the inequality direction.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \leq g^{-1}(y)) \quad \text{(applying } g^{-1} \text{ to both sides)} \\ &= F_X(g^{-1}(y)) \end{aligned}$$

- Decreasing: $\{g(X) \leq y\} \iff \{X \geq g^{-1}(y)\}$. Then, $F_Y(y) = 1 - F_X(g^{-1}(y))$.
  - *Proof:* Since g is decreasing, $u \leq v \iff g(u) \geq g(v)$. Applying g^{-1} reverses the inequality direction.

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) \\ &= \mathbb{P}(X \geq g^{-1}(y)) \quad \text{(inequality flips)} \\ &= 1 - \mathbb{P}(X < g^{-1}(y)) \\ &= 1 - F_X(g^{-1}(y)) \quad \text{(since } X \text{ is continuous, } \mathbb{P}(X = x) = 0) \end{aligned}$$

Theorem (Change of Variables). Suppose X is a continuous RV with PDF f_X, and let $Y = g(X)$. If g is differentiable and strictly monotone (invertible), then:

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

More generally, if g is not invertible but can be partitioned into a countable number of disjoint intervals on which it is strictly monotone, then:

$$f_Y(y) = \sum_{x \in g^{-1}(\{y\})} \frac{f_X(x)}{|g'(x)|}$$

- *Proof:* We differentiate the CDF derived above.
  - Case 1 (g increasing): $f_Y(y) = \frac{d}{dy} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \cdot \frac{d}{dy}(g^{-1}(y))$. The derivative of an inverse is, $\frac{d}{dy}(g^{-1}(y)) = \frac{1}{g'(x)}$. Since g is increasing, $g' > 0$, so this equals $\frac{1}{|g'(x)|}$.
  - Case 2 (g decreasing): $f_Y(y) = \frac{d}{dy}(1 - F_X(g^{-1}(y))) = -f_X(g^{-1}(y)) \cdot \frac{1}{g'(x)}$. Since g is decreasing, $g' < 0$, so $-\frac{1}{g'(x)} = \frac{1}{|g'(x)|}$.

Example. Suppose $X \sim \text{Unif}([0,1])$. Let $\lambda > 0$ and consider $Y = g(X)$ where $g(x) = \frac{-1}{\lambda} \ln(1-x)$. If $X \in (0,1)$, then $Y \in (0, \infty)$. If $b < 0$, then $F_Y(b) = 0$. If $b \geq 0$, then

$$\begin{aligned} F_Y(b) &= \mathbb{P}\left(\frac{-1}{\lambda} \ln(1-X) \leq b\right) \\ &= \mathbb{P}(\ln(1-X) \geq -\lambda b) \\ &= \mathbb{P}(1 - X \geq e^{-\lambda b}) = \mathbb{P}(X \leq 1 - e^{-\lambda b}) \end{aligned}$$

This matches the CDF for an exponential distribution, so by uniqueness, we conclude $Y \sim \text{Exp}(\lambda)$.

6 Joint Distributions

6.1 Discrete RVs

Suppose Ω is the sample space with $X_1...X_n$ being discrete RVs. Thus $\vec{X} = (X_1...X_n)$ is a vector that takes values in $\mathbb{R}^n$.

Definition.

- **Joint Distribution:** $\mathbb{P}(\vec{X} \in B)$ for $B \subseteq \mathbb{R}^n$.
- **Marginal Distribution:** $\mathbb{P}(X_j \in A)$ for $A \subseteq \mathbb{R}$, $j \in \{1, ..., n\}$.
- **Joint PMF:** $\text{pmf}(\vec{k}) = \mathbb{P}(\vec{X} = \vec{k})$ over the possible values of $\vec{X}$.
- **Expectation:** $\mathbb{E}(g(\vec{X})) = \sum_{\vec{k}} g(\vec{k})\mathbb{P}(\vec{X} = \vec{k})$

Remark. One can get the marginal PMFs from the joint PMF by fixing one index and letting all others vary over their possible values (summing $n-1$ times).

6.2 Continuous RVs

Definition (Joint Density). The RVs $X_i...X_k$ are **jointly continuous** if there exists a Joint Density Function (JDF) $f : \mathbb{R}^k \to \mathbb{R}_{\geq 0}$ such that:

$$(\forall\ B \subseteq \mathbb{R}^k) : \mathbb{P}(\vec{X} \in B) = \int_B f(\vec{x}) d^k r$$

Definition (Expectation). Given $g : \mathbb{R}^k \to \mathbb{R}$, $\mathbb{E}(g(\vec{X})) = \int_{\mathbb{R}^k} g(\vec{x}) f(\vec{x}) d^k r$

Remark. A set of continuous RVs $X_1...X_n$ that have well-defined PDFs $f_i(x_i)$ need not have a joint JDF f.

Definition (Marginal CDF). The marginal cdf of X_k without loss of generality is

$$F_{x_k}(a) = \int_{-\infty}^{a} \left(\int_{\mathbb{R}^{k-1}} f(\vec{x}) d^{k-1} r \right) dx_k$$

- By differentiating, the marginal PDF is $f_{x_k}(a) = \int_{\mathbb{R}^{k-1}} f(\vec{x}) d^{k-1} r$.

Definition (Joint Uniform Distribution). We generalize the notion of a uniform distribution over a subset $D \subseteq \mathbb{R}^k$ by a constant jdf inside D and 0 outside.

- Normalization: $f(\vec{x}) = (\int_D d^k r)^{-1}$
- Probability: $\mathbb{P}(\vec{X} \in G) = \frac{\text{n-dim volume of } G \cap D}{\text{n-dim volume of } D}$

6.3 Independent RVs

Proposition. Suppose $X_1...X_n$ are independent discrete RVs. Then

$$\mathbb{P}(\vec{X} = \vec{k}) = \prod_{j=1}^{n} \mathbb{P}(X_j = k_j)$$

- *Proof:* By the definition of mutual independence for random variables, for any events $A_1, \ldots, A_n$ where A_j relates only to X_j, the probability of their intersection is the product of their individual probabilities. Let A_j be the event $\{X_j = k_j\}$ and A the event $\{\vec{X} = \vec{k}\}$. Then $A = \bigcap A_j$ means $\mathbb{P}(\vec{X} = \vec{k}) = \mathbb{P}\left(\bigcap_{j=1}^{n} \{X_j = k_j\}\right) = \prod_{j=1}^{n} \mathbb{P}(X_j = k_j)$.

Proposition. Suppose $X_1...X_k$ are continuous independent RVs with PDFs f_{x_i}. Then

$$f(\vec{x}) = \prod_{i=1}^{k} f_{x_i}(x_i)$$

- *Proof:* Let $F(\vec{x}) = \mathbb{P}(X_1 \leq x_1, \ldots, X_k \leq x_k)$ be the Joint CDF. Independence means the CDF factors into $F(\vec{x}) = \prod_{i=1}^{k} F_{X_i}(x_i)$. We get the joint pdf from the partial derivatives of the joint cdf $f(\vec{x}) = \frac{\partial^k}{\partial x_1 ... \partial x_k} F(\vec{x}) = \frac{\partial^k}{\partial x_1 ... \partial x_k} \left(\prod_{i=1}^{k} F_{X_i}(x_i)\right) = \prod_{i=1}^{k} f_{x_i}(x_i)$ since each $F_{X_i}(x_i)$ depends only on x_i and the partial derivatives apply to each term independently.

Proposition. Suppose $X_1...X_m, Y_1...Y_n$ are independent RVs and that $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$. Defining $F = f(X_1...X_m)$ and $G = g(Y_1...Y_n)$, then F and G are independent RVs.

- *Proof:* To show F and G are independent, we must show that for any sets $A, B \subseteq \mathbb{R}$, $\mathbb{P}(F \in A, G \in B) = \mathbb{P}(F \in A)\mathbb{P}(G \in B)$. Let S_X and S_Y be the pre-images of A and B respectively. The event $\{F \in A\}$ depends only on the variables $\vec{X} = (X_1, \ldots, X_m)$ and similarly for $G \in B$. Since the combined collection $\{X_1, \ldots, X_m, Y_1, \ldots, Y_n\}$ are mutually independent, the vectors $\vec{X}$ and $\vec{Y}$ are independent.

7 Expectation and Variance of Multivariable Distributions

7.1 Linearity of Expectation

Proposition. Given random variables $X_1, \ldots, X_n$ on the sample space, and functions $g_1, \ldots, g_n$, then $\mathbb{E}(\sum_i g_i(X_i)) = \sum_i \mathbb{E}(g_i(X_i))$ In particular, $\mathbb{E}(\sum_i X_i) = \sum_i \mathbb{E}(X_i)$. The only requirement is that the expectations are finite. Otherwise, this is a general argument.

Example. I deal 5 cards with replacement. What is the expected number of aces dealt?

- Let $X = I_1 + \cdots + I_5$ where I_j is an indicator variable such that $I_j = 1$ if the j-th card is an Ace, and 0 otherwise.
- Then $\mathbb{E}(X) = \sum_i \mathbb{E}(I_i)$.
- Since the deals are uniform random with replacement, $\mathbb{P}(I_i = 1) = \frac{4}{52}$.
- $\mathbb{E}(X) = 5 \cdot \frac{4}{52} = \frac{5}{13}$.

Example. Flip a fair coin 100 times. What is the expected number of runs of heads of length 3? (Here we count exact lengths, not subsequences inside bigger lengths.)

- Let I_j be the indicator that a run of length 3 starts at the j-th flip.
- Boundary Cases ($j = 1$ or $j = 98$):
  - For $j = 1$: Sequence must be $HHHT\ldots$ Probability is $(1/2)^4$.
  - For $j = 98$: Sequence must be $\ldots THHH$. Probability is $(1/2)^4$.
- Middle Cases ($1 < j < 98$):
  - Sequence must be $\ldots THHHT\ldots$ Probability is $(1/2)^5$.
- Thus, $\mathbb{E}(X) = \sum_{j=1}^{98} \mathbb{E}(I_j) = 2\left(\frac{1}{16}\right) + 96\left(\frac{1}{32}\right) = \frac{1}{8} + 3 = 3.125$

7.2 Expectation of Independent RVs

Proposition. Given $X_1, \ldots, X_n$ independent RVs and functions $g_1, \ldots, g_n$,

$$\mathbb{E}\left(\prod_{i=1}^{n} g_i(X_i)\right) = \prod_{i=1}^{n} \mathbb{E}(g_i(X_i))$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$$

- *Proof:* We show this for two RVs since we may then apply induction

– For expectation, if X, and Y are discrete then

$$\begin{aligned}\mathbb{E}(g(X)h(Y)) &= \sum_x \sum_y g(x)h(y)\mathbb{P}(X=x, Y=y) \\ &= \sum_x \sum_y g(x)h(y)\mathbb{P}(X=x)\mathbb{P}(Y=y) \quad \text{(by independence)} \\ &= \left(\sum_x g(x)\mathbb{P}(X=x)\right)\left(\sum_y h(y)\mathbb{P}(Y=y)\right) \\ &= \mathbb{E}(g(X))\mathbb{E}(h(Y))\end{aligned}$$

if they are continuous then

$$\begin{aligned}\mathbb{E}(g(X)h(Y)) &= \iint_{\mathbb{R}^2} g(x)h(y)f_{X,Y}(x,y)dxdy \\ &= \iint_{\mathbb{R}^2} g(x)h(y)f_X(x)f_Y(y)dxdy \quad \text{(by independence)} \\ &= \left(\int_{\mathbb{R}} g(x)f_X(x)dx\right)\left(\int_{\mathbb{R}} h(y)f_Y(y)dy\right) \\ &= \mathbb{E}(g(X))\mathbb{E}(h(Y))\end{aligned}$$

and if one is discrete and the other is continuous then we use linearity of sums and integrals to achieve the desired result.

– For variance,

$$\begin{aligned}\text{Var}(X+Y) &= \mathbb{E}((X+Y-\mathbb{E}(X+Y))^2) \\ &= \mathbb{E}(((X-\mu_x)+(Y-\mu_y))^2) \\ &= \mathbb{E}((X-\mu_x)^2) + \mathbb{E}((Y-\mu_y)^2) + 2\mathbb{E}((X-\mu_x)(Y-\mu_y)) \\ &= \sigma^2(X) + \sigma^2(Y) + 2\mathbb{E}(X-\mu_x)\mathbb{E}(Y-\mu_y)\end{aligned}$$

and since X, Y are independent, $\mathbb{E}((X-\mu_x)(Y-\mu_y)) = \mathbb{E}(X-\mu_x)\mathbb{E}(Y-\mu_y) = 0 \cdot 0 = 0$.

7.3 Sample Means and Sample Variances

Proposition. Suppose $X_1, \ldots, X_n$ are RVs that are **independent and identically distributed** (iid). The **sample mean** is $\overline{X}_n = \frac{X_1+\cdots+X_n}{n}$. Then,

- Expectation: $\mathbb{E}(\overline{X}_n) = \mathbb{E}(\frac{1}{n}\sum X_i) = \frac{1}{n}(n\mathbb{E}(X_i)) = \mathbb{E}(X_i)$.
- Variance: $\sigma^2(\overline{X}_n) = \sigma^2(\frac{1}{n}\sum X_i) = \frac{1}{n^2}\sum \sigma^2(X_i) = \frac{n\sigma^2(X_i)}{n^2} = \frac{\sigma^2(X_i)}{n}$.

Proposition. Suppose X and Y are independent RVs. Consider $X+Y$.

$$M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX}e^{tY}) = M_X(t) \cdot M_Y(t)$$

7.4 Covariance and Correlation

Definition. Recall for two RVs, $\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\mathbb{E}((X-\mu_x)(Y-\mu_y))$. We define the last term as the **covariance**. By linearity, $\text{Cov}(X,Y) := \mathbb{E}((X-\mu_x)(Y-\mu_y)) = \mathbb{E}(XY) - \mu_x\mu_y$.

- Note that $(X-\mu_x)(Y-\mu_y) > 0$ if both have the same sign (both above or below mean).
- Note that $(X-\mu_x)(Y-\mu_y) < 0$ if they have different signs.
- Thus, if X and Y tend to move in the same direction, $\text{Cov}(X,Y) > 0$. If they tend in opposite directions, $\text{Cov}(X,Y) < 0$.

Definition. The **correlation** is defined as $\text{Corr}(X,Y) \equiv \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$.

- X, Y are said to be positively correlated if $\text{Corr}(X,Y) > 0$.
- X, Y are said to be negatively correlated if $\text{Corr}(X,Y) < 0$.
- X, Y are said to be uncorrelated if $\text{Corr}(X,Y) = 0$.
- By Cauchy-Schwarz, $-1 \le \text{Corr}(X,Y) \le 1$.
  - *Proof:* Let $\tilde{X} = \frac{X-\mu_X}{\sigma_X}$ and $\tilde{Y} = \frac{Y-\mu_Y}{\sigma_Y}$. Variance is nonnegative meaning

$$0 \le \text{Var}(\tilde{X}+\tilde{Y}) = \text{Var}(\tilde{X}) + \text{Var}(\tilde{Y}) + 2\,\text{Cov}(\tilde{X},\tilde{Y}) = 1+1+2\rho = 2(1+\rho)$$
$$0 \le \text{Var}(\tilde{X}-\tilde{Y}) = \text{Var}(\tilde{X}) + \text{Var}(\tilde{Y}) - 2\,\text{Cov}(\tilde{X},\tilde{Y}) = 1+1-2\rho = 2(1-\rho)$$

Proposition. $\text{Cov}(aX+b, cY+d) = \text{Cov}(aX, cY) = ac \cdot \text{Cov}(X,Y)$

- *Proof:* Using the definition of covariance:

$$\begin{aligned}
\text{Cov}(aX+b, cY+d) &= \mathbb{E}(((aX+b) - \mathbb{E}(aX+b))((cY+d) - \mathbb{E}(cY+d))) \\
&= \mathbb{E}((aX - a\mathbb{E}(X))(cY - c\mathbb{E}(Y))) \quad \text{(constants } b, d \text{ cancel)} \\
&= \mathbb{E}(ac(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) \\
&= ac\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = ac\,\text{Cov}(X,Y)
\end{aligned}$$

Proposition. $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i \ne j} \text{Cov}(X_i, X_j)$

- *Proof:* Let $\mu_i = \mathbb{E}(X_i)$.

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^n X_i\right) &= \mathbb{E}\left(\left(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i\right)^2\right) = \mathbb{E}\left(\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2\right) \\
&= \mathbb{E}\left(\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu_i)(X_j - \mu_j)\right) = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}((X_i - \mu_i)(X_j - \mu_j)) \\
&= \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Cov}(X_i, X_i) + \sum_{i \ne j} \text{Cov}(X_i, X_j)
\end{aligned}$$

7.5 Bivariate Normal Distribution

Definition. Suppose $Z, W \sim N(0,1)$ with $Z \perp W$. Let $\sigma_x, \sigma_y > 0$ and $\rho \in [-1,1]$. We construct X and Y as follows:

$$\begin{aligned} X &\equiv \mu_x + \sigma_x Z && \sim N(\mu_x, \sigma_x^2) \\ Y &\equiv \mu_y + \sigma_y(\rho Z + \sqrt{1-\rho^2}W) && \sim N(\mu_y, \sigma_y^2) \end{aligned}$$

Then,

- Covariance:

$$\begin{aligned} \operatorname{Cov}(X,Y) &= \operatorname{Cov}(\mu_x + \sigma_x Z, \mu_y + \sigma_y(\rho Z + \sqrt{1-\rho^2}W)) \\ &= (\sigma_x\sigma_y)\operatorname{Cov}(Z, \rho Z + \sqrt{1-\rho^2}W) \\ &= \sigma_x\sigma_y(\operatorname{Cov}(Z,\rho Z) + \operatorname{Cov}(Z, \sqrt{1-\rho^2}W)) \\ &= \sigma_x\sigma_y\rho\operatorname{Var}(Z) + 0 = \sigma_x\sigma_y\rho \end{aligned}$$

- Correlation: $\operatorname{Corr}(X,Y) = \frac{\sigma_x\sigma_y\rho}{\sigma_x\sigma_y} = \rho$.
- Joint PDF: $f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right)$
  - *Proof:* We first find the cdf, $F_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y)$. Let $u = \frac{x-\mu_x}{\sigma_x}$ and $v = \frac{y-\mu_y}{\sigma_y}$.

$$\begin{aligned} F_{X,Y}(x,y) &= \mathbb{P}(\mu_x + \sigma_x Z \le x,\ \mu_y + \sigma_y(\rho Z + \sqrt{1-\rho^2}W) \le y) \\ &= \mathbb{P}\left(Z \le u,\ \rho Z + \sqrt{1-\rho^2}W \le v\right) = \mathbb{P}\left(Z \le u,\ W \le \frac{v-\rho Z}{\sqrt{1-\rho^2}}\right) \\ &= \int_{-\infty}^{u}\left(\int_{-\infty}^{\frac{v-\rho z}{\sqrt{1-\rho^2}}} f_W(w)dw\right) f_Z(z)dz = \int_{-\infty}^{u}\Phi\left(\frac{v-\rho z}{\sqrt{1-\rho^2}}\right) f_Z(z)dz. \end{aligned}$$

  To find the joint PDF $f_{X,Y}(x,y)$, we take the mixed partial derivative $\frac{\partial^2}{\partial x\partial y}F_{X,Y}(x,y)$. By FTC, $\frac{\partial F}{\partial x} = \frac{\partial F}{\partial u}\frac{du}{dx} = \left[\Phi\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right) f_Z(u)\right]\cdot\frac{1}{\sigma_x}$ and $\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial x}\right) = \frac{f_Z(u)}{\sigma_x}\cdot\frac{\partial}{\partial v}\Phi\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right)\cdot\frac{dv}{dy}$. But this is $\frac{f_Z(u)}{\sigma_x}\cdot f_W\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right)\frac{1}{\sqrt{1-\rho^2}}\cdot\frac{1}{\sigma_y} = \frac{1}{\sigma_x\sigma_y\sqrt{1-\rho^2}} f_Z(u) f_W\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right)$. Since $Z, W \sim N(0,1)$, $f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2}u^2\right)\exp\left(-\frac{1}{2}\left(\frac{v-\rho u}{\sqrt{1-\rho^2}}\right)^2\right)$ and we may then simplify.

Proposition.

- If $\rho = 0$, then $f_{X,Y}(x,y) = f_X(x)f_Y(y)$, so they are independent.
- If $\rho = \pm 1$, then X and Y are linearly related with $X = \mu_x + \sigma_x Z$ and $Y = \mu_y \pm \sigma_y Z$.
- Joint MGF: $M_{X,Y}(t,s) = \mathbb{E}(e^{tX+sY}) = \exp(t\mu_x + s\mu_y + \frac{1}{2}(\sigma_x^2 t^2 + \sigma_y^2 s^2 + 2\rho\sigma_x\sigma_y ts))$

– *Proof:* Exponentiating and using the linearity of expectation:

$$\begin{aligned}
\mathbb{E}(e^{tX+sY}) &= \mathbb{E}\left(e^{t\mu_x+s\mu_y} \cdot e^{(t\sigma_x+s\sigma_y\rho)Z} \cdot e^{(s\sigma_y\sqrt{1-\rho^2})W}\right) \\
&= e^{t\mu_x+s\mu_y} \cdot \mathbb{E}\left(e^{(t\sigma_x+s\sigma_y\rho)Z}\right) \cdot \mathbb{E}\left(e^{(s\sigma_y\sqrt{1-\rho^2})W}\right) \\
&= e^{t\mu_x+s\mu_y} \cdot \exp\left(\frac{(t\sigma_x+s\sigma_y\rho)^2}{2}\right) \cdot \exp\left(\frac{s^2\sigma_y^2(1-\rho^2)}{2}\right) \\
&= \exp\left(t\mu_x + s\mu_y + \frac{1}{2}\left(t^2\sigma_x^2 + s^2\sigma_y^2\rho^2 + 2ts\sigma_x\sigma_y\rho + s^2\sigma_y^2 - s^2\sigma_y^2\rho^2\right)\right) \\
&= \exp\left(t\mu_x + s\mu_y + \frac{1}{2}\left(t^2\sigma_x^2 + s^2\sigma_y^2 + 2ts\sigma_x\sigma_y\rho\right)\right)
\end{aligned}$$

8 Conditional Distributions

8.1 Conditional Distribution on Discrete RVs

Definition. Suppose B is an event with $\mathbb{P}(B) > 0$ and X is a discrete RV. The **conditional PMF** is defined as:

$$\text{pmf}_{X|B}(k) = \mathbb{P}(X = k|B) = \frac{\mathbb{P}(X = k \cap B)}{\mathbb{P}(B)}$$

Definition (Conditional Expectation). $\mathbb{E}(X|B) = \sum_k k \cdot \text{pmf}_{X|B}(k)$

Proposition (Total Expectation). If $B_1, \ldots, B_n$ form a partition of the sample space Ω, then

- $\text{pmf}_X(k) = \sum_{i=1}^n \text{pmf}_{X|B_i}(k)\mathbb{P}(B_i)$
- $\mathbb{E}(X) = \sum_{i=1}^n \mathbb{E}(X|B_i)\mathbb{P}(B_i)$

Proposition. Suppose X and Y are two RVs. The conditional PMF of X given Y is:

$$\text{pmf}_{X|Y}(k|j) = \mathbb{P}(X = k|Y = j) = \frac{\mathbb{P}(X = k \cap Y = j)}{\mathbb{P}(Y = j)}$$

- Expectation: $\mathbb{E}(X|Y = j) = \sum_k k \cdot \text{pmf}_{X|Y}(k|j)$
- Relationship: $\text{pmf}_{X,Y}(k, j) = \text{pmf}_{X|Y}(k, j)\mathbb{P}(Y = j)$
- Recovery: $\sum_j \text{pmf}_{X|Y}(k|j)\mathbb{P}(Y = j) = \text{pmf}_X(k)$.

8.2 Conditional Distribution of Jointly Continuous RVs

Definition. Given two jointly continuous RVs X, Y, the **conditional density function** is:

$$f_{X|Y}(x|y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} & \text{if } f_Y(y) > 0 \\ 0 & \text{if } f_Y(y) = 0 \end{cases}$$

Proposition. The conditional density function satisfies the following:

- Probability: $\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$
- Expectation: $\mathbb{E}(g(X)|Y = y) = \int_{\mathbb{R}} g(x)f_{X|Y}(x|y)dx$
- Independence: $X \perp Y \iff (\forall x, y)\ f_{X|Y}(x|y) = f_X(x)$
  - *Proof:* $X \perp Y \iff f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_X(x)f_Y(y)}{f_Y(y)} = f_X(x)$
- Total Expectation: $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} \mathbb{E}(g(X)|Y = y)f_Y(y)dy$

– *Proof:*

$$\begin{aligned}\int_{-\infty}^{\infty} \mathbb{E}(g(X)|Y=y)f_Y(y)dy &= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx\right] f_Y(y)dy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)\frac{f_{X,Y}(x,y)}{f_Y(y)}f_Y(y)dxdy \\ &= \iint_{\mathbb{R}^2} g(x)f_{X,Y}(x,y)dxdy \\ &= \mathbb{E}(g(X))\end{aligned}$$

8.3 Conditional Expectation as a Random Variable

Definition. Given two RVs X, Y, we have $v(y) = \mathbb{E}(X|Y=y)$. We define the random variable $\mathbb{E}(X|Y) = v(Y)$ (a function on a RV).

Proposition. The conditional expectation satisfies the following properties:

- Linearity: $\mathbb{E}(aX_1 + X_2|Y) = a\mathbb{E}(X_1|Y) + \mathbb{E}(X_2|Y)$
- Averaging (Tower Property): $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$
  - *Proof:* Let $g(y) = \mathbb{E}(X|Y=y)$. Then $\mathbb{E}(X|Y) = g(Y)$. Then $\mathbb{E}(g(Y)) = \int_{-\infty}^{\infty} g(y)f_Y(y)dy$. But by total expectation, this is $\mathbb{E}(X)$
- Independence: If $X \perp Y$, then $\mathbb{E}(X|Y) = \mathbb{E}(X)$
  - *Proof:* $X \perp Y \Leftrightarrow f_{X|Y}(x|y) = f_X(x) \Leftrightarrow \mathbb{E}(X|Y=y) = \int x f_{X|Y}(x|y)dx = \int x f_X(x)dx = \mathbb{E}(X)$
- Taking out what is known: $\mathbb{E}(f(X)g(Y)|Y) = g(Y)\mathbb{E}(f(X)|Y)$
  - *Proof:* Fix $Y = y$. $\mathbb{E}(f(X)g(Y)|Y=y) = \int (f(x)g(y))f_{X|Y}(x|y)dx = g(y)\int f(x)f_{X|Y}(x|y)dx$ since $g(y)$ is constant wrt x. But y was arbitrary.

Example. Break a stick of length 1 at a random point Y, then break it again at $X < Y$.

- Setup: $Y \sim \text{Unif}(0,1)$ and $X|Y \sim \text{Unif}(0,Y)$.
- Densities: $f_Y(y) = 1$ for $y \in (0,1)$. $f_{X|Y}(x|y) = 1/y$ for $0 < x < y < 1$.
- Marginal PDF of X: $f_X(x) = \int f_{X,Y}dy = \int f_{X|Y}f_Y dy = \int_x^1 \frac{1}{y}dy = \ln(1) - \ln(x) = \ln\left(\frac{1}{x}\right)$
- Expectation: $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(Y/2) = \frac{1}{2}\mathbb{E}(Y) = 1/4$.
- Variance: $\mathbb{E}(X^2) = \mathbb{E}(\mathbb{E}(X^2|Y)) = \mathbb{E}\left(\int_0^y x^2\frac{1}{y}dx\right) = \mathbb{E}\left(\frac{y^2}{3}\right) = \frac{1}{3}\mathbb{E}(Y^2)$. Since $\text{Var}(Y) = 1/12$ and $\mathbb{E}(Y)^2 = 1/4$, $\mathbb{E}(Y^2) = 1/3$. Thus $\mathbb{E}(X^2) = 1/9$ and $\text{Var}(X) = \frac{1}{9} - \left(\frac{1}{4}\right)^2 = \frac{7}{144}$.

Example (Bernoulli Sums). Let $X_1, \ldots$ be iid $\text{Ber}(p)$ RVs. Let $S_k \equiv \sum_{i=1}^k X_i$. Suppose $m < n$, obtain $\mathbb{E}(S_m|S_n)$.

- We calculate $\mathbb{E}(S_m|S_n = k) = \sum_{i=1}^{m} \mathbb{E}(X_i|S_n = k)$ by linearity.
- By iid symmetry, $\mathbb{E}(X_i|S_n = k) = \frac{k}{n}$ (each contributes equally to the sum).
- Probability of a success in the remaining $n-1$ trials is given by

$$\frac{m \cdot p \cdot \mathbb{P}(S_n = k|X_i = 1)}{\mathbb{P}(S_n = k)} = \frac{m \cdot p\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{mk}{n}$$

- Thus, $\mathbb{E}(S_m|S_n) = \frac{mS_n}{n}$.

9 Tail Bounds and Limit Theorems

9.1 Estimating Tail Probabilities

Theorem (Markov's Inequality). Consider a non-negative RV X and constant $c > 0$.

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$$

- *Proof:* Consider the indicator $\mathbb{I}(X \geq c)$. Since $X \geq 0$, ff $X \geq c$, then $X \geq c \cdot 1$, and if $X < c$, then $X \geq c \cdot 0$. Thus, $X \geq c\mathbb{I}(X \geq c)$. Taking expectations, $\mathbb{E}(X) \geq \mathbb{E}(c\mathbb{I}(X \geq c)) = c\mathbb{P}(X \geq c)$.

Theorem (Chebyshev's Inequality). Suppose RV X has finite $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{Var}(X)$. Then, $\mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

- *Proof:* Apply Markov's Inequality to the non-negative RV $(X - \mu)^2$. Namely, $\mathbb{P}(|X - \mu| \geq c) = \mathbb{P}((X - \mu)^2 \geq c^2) \leq \frac{\mathbb{E}((X-\mu)^2)}{c^2} = \frac{\sigma^2}{c^2}$

9.2 Law of Large Numbers

Theorem (Weak Law of Large Numbers). Suppose $X_1, X_2, \ldots$ are iid RVs with $\mu \equiv \mathbb{E}(X_i)$ and $\sigma^2 \equiv \text{Var}(X_i)$ finite. Define $S_n \equiv \sum_{i=1}^n X_i$. Then for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| < \epsilon\right) = 1$

- *Proof:* Recall $\mathbb{E}(S_n/n) = \mu$ and $\text{Var}(S_n/n) = \frac{\sigma^2}{n}$. By Chebyshev's Inequality, $\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\text{Var}(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$. As $n \to \infty$, $\frac{\sigma^2}{n\epsilon^2} \to 0$.

9.3 Central Limit Theorem

Theorem (Central Limit Theorem). Let $X_1, \ldots, X_n$ be iid RVs with mean μ and variance σ^2. Let $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Then as $n \to \infty$, $Z_n \to N(0, 1)$.

- *Proof:*
  - Define $Y_i = \frac{X_i - \mu}{\sigma}$. Then $\mathbb{E}(Y_i) = 0$ and $\text{Var}(Y_i) = 1$
  - Then, $Z_n = \frac{\sum X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum(X_i - \mu)}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i$
  - Let $M_Y(t)$ be the MGF of Y_i. Since Y_i are iid, the MGF of the sum $\sum Y_i$ is $[M_Y(t)]^n$.
  - Since $M_{aX}(t) = M_X(at)$, we have $M_{Z_n}(t) = M_{\frac{1}{\sqrt{n}}\sum Y_i}(t) = M_{\sum Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left[M_Y\left(\frac{t}{\sqrt{n}}\right)\right]^n$
  - Expand $M_Y(s)$ around $s = 0$. Recall that $M(0) = 1$, $M'(0) = \mathbb{E}(Y) = 0$, and $M''(0) = \mathbb{E}(Y^2) = 1$. Thus, $M_Y(s) \approx M_Y(0) + sM_Y'(0) + \frac{s^2}{2}M_Y''(0) = 1 + 0 + \frac{s^2}{2} = 1 + \frac{s^2}{2}$ With $s = \frac{t}{\sqrt{n}}$, $M_Y\left(\frac{t}{\sqrt{n}}\right) \approx 1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 = 1 + \frac{t^2}{2n}$

– Then $\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left(1+\frac{t^2/2}{n}\right)^n = e^{t^2/2}$. We recognize $e^{t^2/2}$ as the MGF of $N(0,1)$, so we conclude that the distributions are equal.