WSABI Summer Research Lab **Summer 2025**

Lecture 1: Review of Probability

Instructor: Jonathan Pipping *Author: JP*

1.1 Introduction

1.1.1 Why Start with Probability?

Probability is the mathematical language we use to describe, quantify, and reason about uncertainty. In any real-world problem – like estimating averages, forecasting stock prices, or classifying images – there is always some inherent randomness or lack of complete information. Specifying a probability model allows us to do the following:

- **Define a data-generating mechanism.** When we say "$X_1, \ldots, X_n$ are i.i.d. samples from some distribution F," we are implicitly modeling how data might vary from one draw to the next. Without a probabilistic description, it's impossible to say "How surprising is this observation?" or "How much confidence do we have in this estimate?"
- **Formulate estimators and quantify their accuracy.** Suppose we want to estimate some parameter θ (e.g. the population mean or variance). A probability model allows us to define estimators $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$, compute their bias and variance, and ultimately derive confidence intervals or credible intervals that tell us how precise our estimate is.
- **Make predictions and assess risk.** In supervised learning, for instance, a probability model yields predictive distributions $P(Y \mid X)$ rather than just point predictions. Knowing the full distribution means we can quantify the risk of making an incorrect decision (e.g. by computing expected loss), set thresholds, or compute prediction intervals that account for noise.
- **Compare competing models via likelihoods.** When we fit two different probability distributions (say, a Normal vs. a skewed Gamma) to the same dataset, we can compare their likelihoods or use information criteria (AIC, BIC) to decide which one explains the data more plausibly. This formal approach to model selection would not be possible without an underlying probabilistic framework.

In short, *probability is the backbone of estimation and inference.* Every time we speak of "confidence intervals," "p-values," "posterior distributions," or "predictive probabilities," we are using probability rules to translate observed data into quantified uncertainty. In the sections that follow, we will review the basic probability rules that we will use to build estimators, hypothesis tests, and predictive models in the coming lectures.

1.1.2 Experiments & Sample Spaces

The first concept we need to understand is that of a random experiment from which we observe certain outcomes. We define this formally as follows:

Definition 1.1 (Random Experiment)**.** *A random experiment is a procedure that can be repeated under identical conditions, for which the outcome cannot be predicted with certainty before it's performed. The set of all possible outcomes of a random experiment is called the **sample space**.*

Definition 1.2 (Sample Space)**.** *The sample space of a random experiment Ω is the set of all possible outcomes of the experiment.*

Here are some examples of random experiments and their sample spaces:

Example 1.3 (Coin Toss)**.** A coin toss is a random experiment with sample space

$$\Omega = \{H, T\},$$

where H represents heads and T represents tails.

Example 1.4 (Dice Roll)**.** A fair dice roll is a random experiment with sample space

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Example 1.5 (Roll of Two Die)**.** Rolling two fair dice is a random experiment with sample space

$$\Omega = \{(i, j) \mid i, j \in \{1, 2, 3, 4, 5, 6\}\}.$$

1.1.3 Events

We now define the concept of an event, which can contain multiple outcomes from the sample space. We also define the power set of a sample space and mutually-exclusive events.

Definition 1.6 (Event)**.** *An event is a subset of the sample space Ω.*

Example 1.7 (Dice-Rolling Events)**.** Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a fair dice roll. The following are all events:

- $A = \{1\}$: The event that the die shows a 1.
- $B = \{2, 4, 6\}$: The event that the die is even.
- $C = \{1, 2, 3, 4, 5, 6\}$: The event that the die shows a number between 1 and 6.
- $D = \emptyset$: The event that the die shows a 7.

Definition 1.8 (The Power Set of Ω)**.** *In finite or countably-infinite settings, the power set of Ω (denoted 2^{Ω}) is the collection of all subsets of Ω (or all events). In more general settings, we work with the set of all* ***measurable subsets*** *of Ω, known as the* ***σ-algebra*** *$\mathcal{F}$.*

We now move on to defining mutually-exclusive and independent events.

1.1.4 Mutual Exclusivity

Definition 1.9 (Mutual Exclusivity)**.** *Two events E_1 and E_2 are* ***mutually exclusive*** *if they cannot both occur at the same time. Formally, $E_1 \cap E_2 = \emptyset$.*

Example 1.10 (Mutually Exclusive Events)**.** Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a fair dice roll. The following are mutually exclusive events:

- $A = \{1\}$: The event that the die shows a 1.
- $B = \{2, 4, 6\}$: The event that the die is even.

Definition 1.11 (Pairwise Mutual Exclusivity)**.** *A collection of events $\{E_i\}_{i=1}^{\infty} = E_1, E_2, \ldots$ is* ***pairwise mutually exclusive*** *if $E_i \cap E_j = \emptyset$ for all $i \neq j$.*

1.1.5 Independence

Definition 1.12 (Independence)**.** *Two events E_1 and E_2 are* ***independent*** *if and only if* $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$.

Example 1.13 (Coin Toss and Dice Roll)**.** Consider a coin toss and a dice roll. What's the probability that the coin lands heads and the dice lands on 1?

$$\begin{aligned}\mathbb{P}(H \cap 1) &= \mathbb{P}(H)\mathbb{P}(1) \text{ by independence}\\ &= \frac{1}{2} \times \frac{1}{6}\\ &= \frac{1}{12}.\end{aligned}$$

With these definitions, we are now ready to formally-define probability.

1.2 Probability

1.2.1 Probability Functions

We begin by defining a probability function, which assigns a probability to each event in the sample space, and some simple porperties.

Definition 1.14 (Probability Function)**.** *A probability function is a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *that satisfies the following axioms:*

- ***Non-negativity:*** $\mathbb{P}(E) \geq 0$ *for all* $E \in \mathcal{F}$.
- ***Normalization:*** $\mathbb{P}(\Omega) = 1$.
- ***Additivity:*** *If* $\{E_i\}_{i=1}^{\infty}$ *is a pairwise mutually exclusive collection of events, then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

Corollary 1.15 (Simple Properties)**.** *If* $\mathbb{P}$ *is a probability function, then*

- $\mathbb{P}(\emptyset) = 0$.
- $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$, *where* E^c *is the complement of event* E.
- $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$ *for any events* E_1 *and* E_2.

Example 1.16 (Probability of a Dice Roll)**.** Consider the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a fair dice roll. The probability function is given by $\mathbb{P}(E) = \frac{|E|}{6}$ for any event $E \in \mathcal{F}$.

Example 1.17 (Rolling Two Dice)**.** Roll two fair dice. Find the probability that the sum is 10 or more.

The sample space Ω consists of all ordered pairs (i, j) where $i, j \in \{1, 2, 3, 4, 5, 6\}$. The event we're interested in is $E = \{(i, j) \mid i + j \geq 10\}$. By counting, we find that

$$E = \{(4, 6), (5, 5), (5, 6), (6, 4), (6, 5), (6, 6)\}.$$

Since each outcome has probability $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$, we have $\mathbb{P}(E) = \frac{6}{36} = \frac{1}{6}$.

We now move on to some essential concepts that will be used throughout the course.

1.2.2 Conditional Probability

We now define the concept of conditional probability, which allows us to compute the probability of an event given that another event has occurred.

Definition 1.18 (Conditional Probability)**.** *Let A and B be events with $\mathbb{P}(B) > 0$. The **conditional probability of** A **given** B is defined as*

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

In other words, once we know that B has occurred, we restrict our "universe" to B (whose probability is $\mathbb{P}(B)$), and then ask: out of the outcomes in B, what fraction also lie in A?

Example 1.19 (Rolling Two Dice)**.** Roll two fair dice. Find the probability of at least one die showing a 6, given that the sum was at least 10.

Let A be the event that at least one die shows a 6. So $A = \{(i,j) \mid i = 6 \text{ or } j = 6\}$. Let B be the event that the sum is at least 10. So $B = \{(i,j) \mid i + j \geq 10\}$. We want to find

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

By counting, we find that $A \cap B = \{(6,4),(4,6),(6,5),(5,6),(6,6)\}$ and $B = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$. And since the probability of each outcome is $\frac{1}{36}$, we have

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{5/36}{6/36} = \frac{5}{6}.$$

Corollary 1.20 (Independence)**.** *We learned in Definition 1.12 that two events are independent if and only if*

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

This is equivalent to saying that

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

Therefore, two events are independent if and only if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ and $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

1.2.3 Law of Total Probability

We now move on to the Law of Total Probability, which allows us to compute the probability of an event directly from its conditional probabilities. First though, we need to define a partition of the sample space.

Definition 1.21 (Partition)**.** *A partition of the sample space Ω is a pairwise mutually exclusive collection of events $\{B_i\}_{i=1}^n$ whose union is the entire sample space Ω. Symbolically,*

$$B_i \cap B_j = \emptyset \quad \textit{for all } i \neq j,$$

and

$$\bigcup_{i=1}^{n} B_i = \Omega.$$

Example 1.22 (Dice Roll)**.** An example of a partition of the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$ is $\{A, A^c\}$, where $A = \{1, 2, 3\}$ and $A^c = \{4, 5, 6\}$.

Theorem 1.23 (Law of Total Probability)**.** *Let $\{B_i\}_{i=1}^n$ be a partition of the sample space Ω. Then for any event $A \in \mathcal{F}$,*

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

In other words, the probability of A can be computed by "breaking" Ω into disjoint pieces B_i, finding the conditional probability of A on each piece, and then averaging those probabilities weighted by $\mathbb{P}(B_i)$

The proof of this Theorem is left as an exercise.

1.2.4 Bayes' Rule

Bayes' Rule is a fundamental theorem in probability that we will use throughout the course. The theorem is as follows:

Theorem 1.24 (Bayes' Rule)**.** *Consider events A and B with $\mathbb{P}(B) > 0$. Then Bayes' Rule says that*

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

And letting $\{A_i\}_{i=1}^n$ be a partition of Ω, we have from the Law of Total Probability that

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).$$

Therefore, we have that

$$\mathbb{P}(A_j \mid B) = \frac{\mathbb{P}(B \mid A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)}.$$

The proof of this Theorem is left as an exercise.

Example 1.25 (Drawing Balls from an Urn)**.** Consider two urns. Urn 1 contains 5 green balls and 7 red balls, while Urn 2 contains 3 green balls and 12 red balls. We flip a fair coin, drawing a ball from Urn 1 if the coin lands heads and from Urn 2 if the coin lands tails. **What is the probability that the coin came up tails given that we drew a green ball?**

We want $\mathbb{P}(T \mid G)$, but much easier is $\mathbb{P}(G \mid T)$. We use Bayes' Rule to simplify this problem.

$$\begin{aligned}
\mathbb{P}(T \mid G) &= \frac{\mathbb{P}(G \mid T)\mathbb{P}(T)}{\mathbb{P}(G)} \text{ by Bayes' Rule} \\
&= \frac{\mathbb{P}(G \mid T)\mathbb{P}(T)}{\mathbb{P}(G \mid H)\mathbb{P}(H) + \mathbb{P}(G \mid T)\mathbb{P}(T)} \text{ by LoTP} \\
&= \frac{(3/15)\,(1/2)}{(5/12)\,(1/2) + (3/15)\,(1/2)} \\
&= 0.324
\end{aligned}$$

We are now ready to move on to the next section, where we will discuss random variables and their distributions.

1.3 Random Variables

1.3.1 What is a Random Variable?

Definition 1.26 (Random Variable)**.** *A random variable X is a real-valued function from the sample space Ω to the real numbers $\mathbb{R}$. Mathematically,*

$$X : \Omega \to \mathbb{R}.$$

Example 1.27 (Rolling Two Dice)**.** Roll two fair dice, and let X be the sum of the two dice. If (i, j) is the outcome of the roll, then $X = i + j$ maps the 36 (i, j) outcomes in the sample space to the real values $\{2, 3, \ldots, 12\}$.

1.3.2 Discrete Random Variables

Definition 1.28 (Discrete Random Variable)**.** *A random variable X is* ***discrete*** *if it takes on a finite or countably infinite number of values.*

This allows us to define the probability mass function (PMF) of X as follows:

Definition 1.29 (Probability Mass Function)**.** *The probability mass function (PMF) of a discrete random variable X is the function $p_X : \mathbb{R} \to [0, 1]$ that satisfies*

$$p_X(x) = \mathbb{P}(X = x).$$

In other words, the PMF of X is the function that gives the probability that X takes on a particular value x.

Properties of the PMF. Let X be a discrete random variable with support S_X. Then

- $p_X(x) \geq 0$ for all $x \in S_X$.
- $\sum_{x \in S_X} p_X(x) = 1$.

Example 1.30 (Rolling Two Dice)**.** Let X be the sum of two fair dice. Then the PMF of X is given by

$$p_X(x) = \begin{cases} \frac{1}{36} & \text{if } x = 2, \\ \frac{2}{36} & \text{if } x = 3, \\ \frac{3}{36} & \text{if } x = 4, \\ \frac{4}{36} & \text{if } x = 5, \\ \frac{5}{36} & \text{if } x = 6, \\ \frac{6}{36} & \text{if } x = 7, \\ \frac{5}{36} & \text{if } x = 8, \\ \frac{4}{36} & \text{if } x = 9, \\ \frac{3}{36} & \text{if } x = 10, \\ \frac{2}{36} & \text{if } x = 11, \\ \frac{1}{36} & \text{if } x = 12 \end{cases}$$

We now list some commonly-used discrete random variables and their PMFs.

Definition 1.31 (Bernoulli Distribution). *A random variable X is said to have a Bernoulli distribution with parameter p if it takes on the value 1 with probability p and the value 0 with probability $1-p$. The PMF of a Bernoulli random variable is given by*

$$p_X(x) = p^x(1-p)^{1-x},\ p \in [0,1],\ x \in \{0,1\}$$

X represents whether a Bernoulli trial is successful or not.

Definition 1.32 (Binomial Distribution). *A random variable X is said to have a binomial distribution with parameters n and p if it takes on the value x with probability*

$$p_X(x) = \binom{n}{x}p^x(1-p)^{n-x},\ p \in [0,1],\ x \in \{0,1,\ldots,n\}$$

X is the sum of n independent Bernoulli(p) random variables, and represents the number of successes in these n trials.

Definition 1.33 (Geometric Distribution). *A random variable X is said to have a geometric distribution with parameter p if it takes on the value x with probability*

$$p_X(x) = p(1-p)^{x-1},\ p \in [0,1],\ x \in \{1,2,\ldots\}$$

X represents the number of Bernoulli trials until the first success.

Definition 1.34 (Poisson Distribution). *A random variable X is said to have a Poisson distribution with parameter λ if it takes on the value x with probability*

$$p_X(x) = \frac{\lambda^x e^{-\lambda}}{x!},\ \lambda > 0,\ x \in \{0,1,\ldots\}$$

X represents the number of events that occur in a fixed interval of time or space, given a known average rate of occurrence.

We will now move on to continuous random variables, which are similar but have a different set of properties.

1.3.3 Continuous Random Variables

Definition 1.35 (Continuous Random Variable). *A random variable X is* ***continuous*** *if it takes on a continuous range of values rather than a discrete set of values.*

Since continuous random variables can take on any value in a continuous range, there are an infinite number of possible realizations. This means that the probability of realizing any particular value is 0, and we cannot define a PMF. Instead, we define the cumulative distribution function (CDF) of X as follows:

Definition 1.36 (Cumulative Distribution Function). *The cumulative distribution function (CDF) of a continuous random variable X is the function $F_X : \mathbb{R} \to [0,1]$ that satisfies*

$$F_X(x) = \mathbb{P}(X \leq x),\ x \in \mathbb{R}$$

Then for any $a, b \in \mathbb{R}$ with $a \leq b$, we have that

$$\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$$

This allows us to define the probability density function (PDF) of X as follows:

Definition 1.37 (Probability Density Function). *The probability density function (PDF) of a continuous random variable X is the function $f_X : \mathbb{R} \to [0,1]$ that satisfies*

$$f_X(x) = \frac{d}{dx} F_X(x),\ x \in \mathbb{R}$$

Then for any $a, b \in \mathbb{R}$ with $a \leq b$, we have that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$$

We now list some commonly-used continuous random variables and their PDFs.

Definition 1.38 (Uniform Distribution). *A random variable X is said to have a uniform distribution on the interval $[a, b]$ if it has the PDF*

$$f_X(x) = \frac{1}{b-a},\ x \in [a,b].$$

Definition 1.39 (Normal Distribution). *A random variable X is said to have a normal distribution with mean μ and variance σ^2 if it has the PDF*

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \mu \in \mathbb{R},\ \sigma > 0,\ x \in \mathbb{R}$$

Definition 1.40 (Exponential Distribution). *A random variable X is said to have an exponential distribution with parameter λ if it has the PDF*

$$f_X(x) = \lambda e^{-\lambda x},\ \lambda > 0,\ x \in [0, \infty)$$

Definition 1.41 (Gamma Distribution). *A random variable X is said to have a gamma distribution with parameters α and β if it has the PDF*

$$f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}},\ \alpha,\ \beta > 0,\ x \in [0, \infty)$$

Definition 1.42 (Beta Distribution). *A random variable X is said to have a beta distribution with parameters α and β if it has the PDF*

$$f_X(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},\ \alpha,\ \beta > 0,\ x \in [0,1]$$

where $B(\alpha, \beta)$ is the beta function

$$\begin{aligned} B(\alpha, \beta) &= \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \\ &= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\ &= \frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!} \end{aligned}$$

We now move on to consider jointly and conditionally distributed random variables.

1.3.4 Joint and Conditional Distributions

We define the joint distributions of two random variables X and Y as follows:

Definition 1.43 (Joint PMF for Discrete RVs)**.** *The joint probability mass function (PMF) of two discrete random variables X and Y is the function $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ that satisfies*

$$f_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$$

Note that we can get the marginal PMF of X by summing the joint PMF over all possible values of Y:

$$f_X(x) = \sum_{y \in S_Y} f_{X,Y}(x,y)$$

where S_Y is the support of Y.

Definition 1.44 (Joint CDF for Continuous RVs)**.** *The joint cumulative distribution function (CDF) of two continuous random variables X and Y is the function $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ that satisfies*

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

Definition 1.45 (Joint PDF for Continuous RVs)**.** *The joint probability density function (PDF) of two continuous random variables X and Y is the function $f_{X,Y} : \mathbb{R}^2 \to [0,1]$ that satisfies*

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

where $F_{X,Y}$ is the joint CDF of X and Y. Then for any $a, b, c, d \in \mathbb{R}$ with $a \leq b$ and $c \leq d$, we have that

$$\mathbb{P}(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y) dx dy$$

Note that we can get the marginal PDF of X by integrating the joint PDF over all possible values of Y:

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

where S_Y is the support of Y.

We now define the conditional distributions of two random variables X and Y as follows:

Definition 1.46 (Conditional PMF)**.** *The conditional probability mass function (PMF) of two discrete random variables X and Y given $Y = y$ is the function $f_{X,Y}(x \mid y) : \mathbb{R} \to [0,1]$ that satisfies*

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

Definition 1.47 (Conditional PDF)**.** *The conditional probability density function (PDF) of two continuous random variables X and Y given $Y = y$ is the function $f_{X,Y}(x \mid y) : \mathbb{R} \to [0,1]$ that satisfies*

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

We now move on to the next section, where we will discuss expectation and variance.

1.4 Expectation and Variance

1.4.1 Expectation

We begin by defining the expectation of a random variable.

Definition 1.48 (Expectation)**.** *The expectation of a random variable X is the average value of X over all possible outcomes. If X is discrete with support S_X, then*

$$\mathbb{E}[X] = \sum_{x \in S_X} x p_X(x)$$

If X is continuous with support S_X, then

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

The expectation is a measure of the center of the distribution of X.

Key Properties of Expectation.

- **Linearity:** $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ for any constants a and b.
- **Additivity:** $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for any random variables X and Y.
- **Independence:** $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ for any independent random variables X and Y.

Definition 1.49 (Expectation of a Function of a Random Variable)**.** *The expectation of a function $g : \mathbb{R} \to \mathbb{R}$ of a random variable X is the average value of the function over all possible outcomes. If X is discrete with support S_X, then*

$$\mathbb{E}[g(X)] = \sum_{x \in S_X} g(x) p_X(x)$$

If X is continuous with support S_X, then

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

Definition 1.50 (Expectation of a Function of Jointly-Distributed Random Variables)**.** *The expectation of a function $g : \mathbb{R}^2 \to \mathbb{R}$ of two random variables X and Y is the average value of the function over all possible outcomes. If X and Y are discrete with support S_X and S_Y, then*

$$\mathbb{E}[g(X, Y)] = \sum_{x \in S_X, y \in S_Y} g(x, y) p_{X,Y}(x, y)$$

If X and Y are continuous with support S_X and S_Y, then

$$\mathbb{E}[g(X, Y)] = \int_{S_Y} \int_{S_X} g(x, y) f_{X,Y}(x, y) dx dy$$

Definition 1.51 (Conditional Expectation)**.** *The conditional expectation of a random variable X given fixed $Y = y$ is the average value of X given $Y = y$. If X and Y are discrete with support S_X and S_Y, then*

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in S_X} x f_{X|Y}(x \mid y)$$

If X and Y are continuous with support S_X and S_Y, then

$$\mathbb{E}[X \mid Y = y] = \int_{S_X} x f_{X|Y}(x \mid y) dx$$

If instead we allow Y to be random, we can define the conditional expectation of X given Y as follows:

Definition 1.52 (Conditional Expectation as a Random Variable)**.** *If we do not fix Y to be a particular value, then $\mathbb{E}[X \mid Y]$ is a random variable **which depends on Y**.*

Theorem 1.53 (Law of Iterated Expectation)**.** *The law of iterated expectation states that*

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]$$

Proof. In the discrete case (without loss of generality), we have that

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in S_X} x\mathbb{P}(X = x) \\
&= \sum_{x \in S_X} x \sum_{y \in S_Y} \mathbb{P}(X = x, Y = y) \\
&= \sum_{x \in S_X} x \sum_{y \in S_Y} \mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} x\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) \\
&= \sum_{y \in S_Y} \sum_{x \in S_X} x\mathbb{P}(X = x \mid Y = y)\mathbb{P}(Y = y) \\
&= \sum_{y \in S_Y} \mathbb{E}[X \mid Y = y]\mathbb{P}(Y = y) \\
&= \mathbb{E}[\mathbb{E}[X \mid Y]]
\end{aligned}$$

□

We are now ready to define the variance of a random variable.

1.4.2 Variance

Definition 1.54 (Variance)**.** *The variance of a random variable X is the expected value of the squared deviation from the mean $\mu = \mathbb{E}[X]$.*

$$\begin{aligned}
Var(X) &= \mathbb{E}[(X - \mu)^2] \\
&= \mathbb{E}[X^2] - \mu^2
\end{aligned}$$

If X is discrete with support S_X, then

$$Var(X) = \mathbb{E}[(X - \mu)^2] = \sum_{x \in S_X} (x - \mu)^2 p_X(x)$$

If X is continuous with support S_X, then

$$Var(X) = \mathbb{E}[(X - \mu)^2] = \int_{S_X} (x - \mu)^2 f_X(x)dx$$

The variance is a measure of the variability of X around its mean.

Definition 1.55 (Covariance)**.** *The covariance of two random variables X and Y is the expected value of the product of the deviations from their means $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.*

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

If X and Y are discrete with support S_X and S_Y, then

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \sum_{x \in S_X, y \in S_Y} (x - \mu_X)(y - \mu_Y) p_{X,Y}(x, y)$$

where $p_{X,Y}(x, y)$ is the joint PMF of X and Y.

If X and Y are continuous with support S_X and S_Y, then

$$Cov(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \int_{S_X} \int_{S_Y} (x - \mu_X)(y - \mu_Y) f_{X,Y}(x, y) dxdy$$

where $f_{X,Y}(x, y)$ is the joint PDF of X and Y.

The covariance is a measure of the level of linear dependence between X and Y.

Theorem 1.56 (Independence Implies Zero Covariance)**.** *If X and Y are independent, then $Cov(X, Y) = 0$.*

Proof.

$$\begin{aligned} \text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] \text{ by independence} \\ &= 0 \end{aligned}$$

□

Key Properties of Variance.

- **Linearity:** $\text{Var}(aX + b) = a^2\text{Var}(X)$ for any constants a and b.

- **Additivity:** $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ for any random variables X and Y.

  Note that from Theorem 1.56, if $X \perp\!\!\!\perp Y$, then $Var(X + Y) = Var(X) + Var(Y)$.

Now, we'll finish up with the expectation and variance of some common distributions.

1.4.3 Expectations & Variances of Common Distributional Families

Distribution X	$\mathbb{E}[X]$	$\mathrm{Var}(X)$
Bernoulli(p)	p	$p(1-p)$
Binomial(n, p)	np	$np(1-p)$
Geometric(p)*	$1/p$	$(1-p)/p^2$
Poisson(λ)	λ	λ
Uniform$[a, b]$	$(a+b)/2$	$(b-a)^2/12$
Normal(μ, σ^2)	μ	σ^2
Exponential(λ)	$1/\lambda$	$1/\lambda^2$
Gamma(α, β)	α/β	α/β^2
Beta(α, β)	$\alpha/(\alpha+\beta)$	$\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$

*Number of trials until the first success (starts at 1). If you define the geometric r.v. as ”counting failures,” just subtract 1 from the mean and leave the variance unchanged.