

# Probability & Statistics with Applications to Computing

## Key Definitions and Theorems

### 1 Combinatorial Theory

#### 1.1 So You Think You Can Count?

**The Sum Rule:** If an experiment can either end up being one of  $N$  outcomes, or one of  $M$  outcomes (where there is no overlap), then the total number of possible outcomes is:  $N + M$ .

**The Product Rule:** If an experiment has  $N_1$  outcomes for the first stage,  $N_2$  outcomes for the second stage, ..., and  $N_m$  outcomes for the  $m^{\text{th}}$  stage, then the total number of outcomes of the experiment is  $N_1 \times N_2 \cdots N_m = \prod_{i=1}^m N_i$ .

**Permutation:** The number of orderings of  $N$  **distinct** objects is  $N! = N \cdot (N - 1) \cdot (N - 2) \cdots 3 \cdot 2 \cdot 1$ .

**Complementary Counting:** Let  $\mathcal{U}$  be a (finite) universal set, and  $S$  a subset of interest. Then,  $|S| = |\mathcal{U}| - |\mathcal{U} \setminus S|$ .

#### 1.2 More Counting

**$k$ -Permutations:** If we want to *pick* (**order matters**) only  $k$  out of  $n$  distinct objects, the number of ways to do so is:

$$P(n, k) = n \cdot (n - 1) \cdot (n - 2) \cdots (n - k + 1) = \frac{n!}{(n-k)!}$$

**$k$ -Combinations/Binomial Coefficients:** If we want to *choose* (**order doesn't matter**) only  $k$  out of  $n$  distinct objects, the number of ways to do so is:

$$C(n, k) = \binom{n}{k} = \frac{P(n, k)}{k!} = \frac{n!}{k!(n-k)!}$$

**Multinomial Coefficients:** If we have  $k$  distinct types of objects ( $n$  total), with  $n_1$  of the first type,  $n_2$  of the second, ..., and  $n_k$  of the  $k$ -th, then the number of arrangements possible is

$$\binom{n}{n_1, n_2, \dots, n_k} = \frac{n!}{n_1! n_2! \dots n_k!}$$

**Stars and Bars/Divider Method:** The number of ways to distribute  $n$  indistinguishable balls into  $k$  distinguishable bins is

$$\binom{n + (k - 1)}{k - 1} = \binom{n + (k - 1)}{n}$$

#### 1.3 No More Counting Please

**Binomial Theorem:** Let  $x, y \in \mathbb{R}$  and  $n \in \mathbb{N}$  a positive integer. Then:  $(x + y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$ .

**Principle of Inclusion-Exclusion (PIE):**

2 events:  $|A \cup B| = |A| + |B| - |A \cap B|$

3 events:  $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$

$k$  events: singles - doubles + triples - quads + ...

**Pigeonhole Principle:** If there are  $n$  pigeons we want to put into  $k$  holes (where  $n > k$ ), then at least one pigeonhole must contain at least 2 (or to be precise,  $\lceil n/k \rceil$ ) pigeons.

**Combinatorial Proofs:** To prove two quantities are equal, you can come up with a combinatorial situation, and show that both in fact count the same thing, and hence must be equal.

### 2 Discrete Probability

#### 2.1 Discrete Probability

**Key Probability Definitions:** The **sample space** is the set  $\Omega$  of all possible outcomes of an experiment. An **event** is any subset  $E \subseteq \Omega$ . Events  $E$  and  $F$  are **mutually exclusive** if  $E \cap F = \emptyset$ .

**Axioms of Probability & Consequences:**

1. (**Axiom: Nonnegativity**) For any event  $E$ ,  $\mathbb{P}(E) \geq 0$ .

2. **(Axiom: Normalization)**  $\mathbb{P}(\Omega) = 1$ .
3. **(Axiom: Countable Additivity)** If  $E$  and  $F$  are mutually exclusive, then  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ .
1. **(Corollary: Complementation)**  $\mathbb{P}(E^C) = 1 - \mathbb{P}(E)$
2. **(Corollary: Monotonicity)** If  $E \subseteq F$ , then  $\mathbb{P}(E) \leq \mathbb{P}(F)$
3. **(Corollary: Inclusion-Exclusion)**  $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

**Equally Likely Outcomes:** If  $\Omega$  is a sample space such that each of the unique outcome elements in  $\Omega$  are equally likely, then for any event  $E \subseteq \Omega$ :  $\mathbb{P}(E) = |E|/|\Omega|$ .

## 2.2 Conditional Probability

**Conditional Probability:**  $\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

**Bayes Theorem:**  $\mathbb{P}(A | B) = \frac{\mathbb{P}(B | A) \mathbb{P}(A)}{\mathbb{P}(B)}$

**Partition:** Non-empty events  $E_1, \dots, E_n$  **partition** the sample space  $\Omega$  if they are both:

- **(Exhaustive)**  $E_1 \cup E_2 \cup \dots \cup E_n = \bigcup_{i=1}^n E_i = \Omega$  (they cover the entire sample space).
- **(Pairwise Mutually Exclusive)** For all  $i \neq j$ ,  $E_i \cap E_j = \emptyset$  (none of them overlap)

Note that for any event  $E$ ,  $E$  and  $E^C$  always form a partition of  $\Omega$ .

**Law of Total Probability (LTP):** If events  $E_1, \dots, E_n$  partition  $\Omega$ , then for any event  $F$ :

$$\mathbb{P}(F) = \sum_{i=1}^n \mathbb{P}(F \cap E_i) = \sum_{i=1}^n \mathbb{P}(F | E_i) \mathbb{P}(E_i)$$

**Bayes Theorem with LTP:** Let events  $E_1, \dots, E_n$  partition the sample space  $\Omega$ , and let  $F$  be another event. Then:

$$\mathbb{P}(E_1 | F) = \frac{\mathbb{P}(F | E_1) \mathbb{P}(E_1)}{\sum_{i=1}^n \mathbb{P}(F | E_i) \mathbb{P}(E_i)}$$

## 2.3 Independence

**Chain Rule:** Let  $A_1, \dots, A_n$  be events with nonzero probabilities. Then:

$$\mathbb{P}(A_1, \dots, A_n) = \mathbb{P}(A_1) \mathbb{P}(A_2 | A_1) \mathbb{P}(A_3 | A_1 A_2) \dots \mathbb{P}(A_n | A_1, \dots, A_{n-1})$$

**Independence:**  $A$  and  $B$  are **independent** if any of the following equivalent statements hold:

1.  $\mathbb{P}(A | B) = \mathbb{P}(A)$
2.  $\mathbb{P}(B | A) = \mathbb{P}(B)$
3.  $\mathbb{P}(A, B) = \mathbb{P}(A) \mathbb{P}(B)$

**Mutual Independence:** We say  $n$  events  $A_1, A_2, \dots, A_n$  are **(mutually) independent** if, for *any* subset  $I \subseteq [n] = \{1, 2, \dots, n\}$ , we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$$

This equation is actually representing  $2^n$  equations since there are  $2^n$  subsets of  $[n]$ .

**Conditional Independence:**  $A$  and  $B$  are **conditionally independent given an event  $C$**  if any of the following equivalent statements hold:

1.  $\mathbb{P}(A | B, C) = \mathbb{P}(A | C)$

2.  $\mathbb{P}(B \mid A, C) = \mathbb{P}(B \mid C)$
3.  $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$

### 3 Discrete Random Variables

#### 3.1 Discrete Random Variables Basics

**Random Variable (RV):** A random variable (RV)  $X$  is a numeric function of the outcome  $X : \Omega \rightarrow \mathbb{R}$ . The set of possible values  $X$  can take on is its **range/support**, denoted  $\Omega_X$ .

If  $\Omega_X$  is finite or countable infinite (typically integers or a subset),  $X$  is a **discrete RV**. Else if  $\Omega_X$  is uncountably large (the size of real numbers),  $X$  is a **continuous RV**.

**Probability Mass Function (PMF):** For a discrete RV  $X$ , assigns probabilities to values in its range. That is  $p_X : \Omega_X \rightarrow [0, 1]$  where:  $p_X(k) = \mathbb{P}(X = k)$ .

**Expectation:** The **expectation** of a discrete RV  $X$  is:  $\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k)$ .

#### 3.2 More on Expectation

**Linearity of Expectation (LoE):** For any random variables  $X, Y$  (possibly dependent):

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

**Law of the Unconscious Statistician (LOTUS):** For a discrete RV  $X$  and function  $g$ ,  $\mathbb{E}[g(X)] = \sum_{b \in \Omega_X} g(b) \cdot p_X(b)$ .

#### 3.3 Variance

**Linearity of Expectation with Indicators:** If asked only about the expectation of a RV  $X$  which is some sort of “count” (and not its PMF), then you may be able to write  $X$  as the sum of possibly dependent **indicator** RVs  $X_1, \dots, X_n$ , and apply LoE, where for an indicator RV  $X_i$ ,  $\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1)$ .

**Variance:**  $\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ .

**Standard Deviation (SD):**  $\sigma_X = \sqrt{\text{Var}(X)}$ .

**Property of Variance:**  $\text{Var}(aX + b) = a^2\text{Var}(X)$ .

#### 3.4 Zoo of Discrete Random Variables Part I

**Independence:** Random variables  $X$  and  $Y$  are **independent**, denoted  $X \perp Y$ , if for *all*  $x \in \Omega_X$  and all  $y \in \Omega_Y$ :  $\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$ .

**Independent and Identically Distributed (iid):** We say  $X_1, \dots, X_n$  are said to be **independent and identically distributed (iid)** if all the  $X_i$ 's are independent of each other, and have the same distribution (PMF for discrete RVs, or CDF for continuous RVs).

**Variance Adds for Independent RVs:** If  $X \perp Y$ , then  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

**Bernoulli Process:** A **Bernoulli process** with parameter  $p$  is a sequence of independent coin flips  $X_1, X_2, X_3, \dots$  where  $\mathbb{P}(\text{head}) = p$ . If flip  $i$  is heads, then we encode  $X_i = 1$ ; otherwise,  $X_i = 0$ .

**Bernoulli/Indicator Random Variable:**  $X \sim \text{Bernoulli}(p)$  ( $\text{Ber}(p)$  for short) iff  $X$  has PMF:

$$p_X(k) = \begin{cases} p, & k = 1 \\ 1 - p, & k = 0 \end{cases}$$

$\mathbb{E}[X] = p$  and  $\text{Var}(X) = p(1 - p)$ . An example of a Bernoulli/indicator RV is one flip of a coin with  $\mathbb{P}(\text{head}) = p$ . By a clever trick, we can write

$$p_X(k) = p^k (1 - p)^{1-k}, \quad k = 0, 1$$

**Binomial Random Variable:**  $X \sim \text{Binomial}(n, p)$  ( $\text{Bin}(n, p)$  for short) iff  $X$  has PMF

$$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in \Omega_X = \{0, 1, \dots, n\}$$

$\mathbb{E}[X] = np$  and  $\text{Var}(X) = np(1 - p)$ .  $X$  is the sum of  $n$  iid  $\text{Ber}(p)$  random variables. An example of a Binomial RV is the number of heads in  $n$  independent flips of a coin with  $\mathbb{P}(\text{head}) = p$ . Note that  $\text{Bin}(1, p) \equiv \text{Ber}(p)$ . As  $n \rightarrow \infty$  and  $p \rightarrow$

0, with  $np = \lambda$ , then  $\text{Bin}(n, p) \rightarrow \text{Poi}(\lambda)$ . If  $X_1, \dots, X_n$  are independent Binomial RV's, where  $X_i \sim \text{Bin}(N_i, p)$ , then  $X = X_1 + \dots + X_n \sim \text{Bin}(N_1 + \dots + N_n, p)$ .

### 3.5 Zoo of Discrete Random Variables Part II

**Uniform Random Variable (Discrete):**  $X \sim \text{Uniform}(a, b)$  ( $\text{Unif}(a, b)$  for short), for integers  $a \leq b$ , iff  $X$  has PMF:

$$p_X(k) = \frac{1}{b - a + 1}, \quad k \in \Omega_X = \{a, a + 1, \dots, b\}$$

$\mathbb{E}[X] = \frac{a+b}{2}$  and  $\text{Var}(X) = \frac{(b-a)(b-a+1)}{12}$ . This represents each *integer* in  $[a, b]$  to be equally likely. For example, a single roll of a fair die is  $\text{Unif}(1, 6)$ .

**Geometric Random Variable:**  $X \sim \text{Geometric}(p)$  ( $\text{Geo}(p)$  for short) iff  $X$  has PMF:

$$p_X(k) = (1 - p)^{k-1} p, \quad k \in \Omega_X = \{1, 2, 3, \dots\}$$

$\mathbb{E}[X] = \frac{1}{p}$  and  $\text{Var}(X) = \frac{1-p}{p^2}$ . An example of a Geometric RV is the number of independent coin flips up to and including the first head, where  $\mathbb{P}(\text{head}) = p$ .

**Negative Binomial Random Variable:**  $X \sim \text{NegativeBinomial}(r, p)$  ( $\text{NegBin}(r, p)$  for short) iff  $X$  has PMF:

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k \in \Omega_X = \{r, r+1, r+2, \dots\}$$

$\mathbb{E}[X] = \frac{r}{p}$  and  $\text{Var}(X) = \frac{r(1-p)}{p^2}$ .  $X$  is the sum of  $r$  iid  $\text{Geo}(p)$  random variables. An example of a Negative Binomial RV is the number of independent coin flips up to and including the  $r$ -th head, where  $\mathbb{P}(\text{head}) = p$ . If  $X_1, \dots, X_n$  are independent Negative Binomial RV's, where  $X_i \sim \text{NegBin}(r_i, p)$ , then  $X = X_1 + \dots + X_n \sim \text{NegBin}(r_1 + \dots + r_n, p)$ .

### 3.6 Zoo of Discrete Random Variables Part III

**Poisson Random Variable:**  $X \sim \text{Poisson}(\lambda)$  ( $\text{Poi}(\lambda)$  for short) iff  $X$  has PMF:

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \Omega_X = \{0, 1, 2, \dots\}$$

$\mathbb{E}[X] = \lambda$  and  $\text{Var}(X) = \lambda$ . An example of a Poisson RV is the number of people born during a particular minute, where  $\lambda$  is the average birth rate per minute. If  $X_1, \dots, X_n$  are independent Poisson RV's, where  $X_i \sim \text{Poi}(\lambda_i)$ , then  $X = X_1 + \dots + X_n \sim \text{Poi}(\lambda_1 + \dots + \lambda_n)$ .

**Hypergeometric Random Variable:**  $X \sim \text{HyperGeometric}(N, K, n)$  ( $\text{HypGeo}(N, K, n)$  for short) iff  $X$  has PMF:

$$p_X(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}, \quad k \in \Omega_X = \{\max\{0, n + K - N\}, \dots, \min\{K, n\}\}$$

$\mathbb{E}[X] = n \frac{K}{N}$  and  $\text{Var}(X) = n \frac{K(N-K)(N-n)}{N^2(N-1)}$ . This represents the number of successes drawn, when  $n$  items are drawn from a bag with  $N$  items ( $K$  of which are successes, and  $N - K$  failures) *without* replacement. If we did this with replacement, then this scenario would be represented as  $\text{Bin}(n, \frac{K}{N})$ .

## 4 Continuous Random Variables

### 4.1 Continuous Random Variables Basics

**Probability Density Function (PDF):** The **probability density function (PDF)** of a continuous RV  $X$  is the function  $f_X : \mathbb{R} \rightarrow \mathbb{R}$ , such that the following properties hold:

- $f_X(z) \geq 0$  for all  $z \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_X(t) dt = 1$
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(w) dw$

**Cumulative Distribution Function (CDF):** The **cumulative distribution function (CDF)** of ANY random variable (discrete or continuous) is defined to be the function  $F_X : \mathbb{R} \rightarrow \mathbb{R}$  with  $F_X(t) = \mathbb{P}(X \leq t)$ . If  $X$  is a *continuous* RV, we have:

- $F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^t f_X(w) dw$  for all  $t \in \mathbb{R}$
- $\frac{d}{du} F_X(u) = f_X(u)$

**Univariate: Discrete to Continuous:**

	Discrete	Continuous
PMF/PDF	$p_X(x) = \mathbb{P}(X = x)$	$f_X(x) \neq \mathbb{P}(X = x) = 0$
CDF	$F_X(x) = \sum_{t \leq x} p_X(t)$	$F_X(x) = \int_{-\infty}^x f_X(t) dt$
Normalization	$\sum_x p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x) dx = 1$
Expectation/LOTUS	$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x) dx$

**4.2 Zoo of Continuous RVs**

**Uniform Random Variable (Continuous):**  $X \sim \text{Uniform}(a, b)$  (Unif( $a, b$ ) for short) iff  $X$  has PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in \Omega_X = [a, b] \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{a+b}{2}$  and  $\text{Var}(X) = \frac{(b-a)^2}{12}$ . This represents each real number from  $[a, b]$  to be equally likely. Do NOT confuse this with its discrete counterpart!

**Exponential Random Variable:**  $X \sim \text{Exponential}(\lambda)$  (Exp( $\lambda$ ) for short) iff  $X$  has PDF:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \in \Omega_X = [0, \infty) \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{1}{\lambda}$  and  $\text{Var}(X) = \frac{1}{\lambda^2}$ .  $F_X(x) = 1 - e^{-\lambda x}$  for  $x \geq 0$ . The exponential RV is the continuous analog of the geometric RV: it represents the waiting time to the next event, where  $\lambda > 0$  is the average number of events per unit time. Note that the exponential measures how much time passes until the next event (any real number, continuous), whereas the Poisson measures how many events occur in a unit of time (nonnegative integer, discrete). The exponential RV is also memoryless:

$$\text{for any } s, t \geq 0, \mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$$

**4.3 The Normal/Gaussian Random Variable**

**Normal (Gaussian, “bell curve”) Random Variable:**  $X \sim \mathcal{N}(\mu, \sigma^2)$  iff  $X$  has PDF:

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

$\mathbb{E}[X] = \mu$  and  $\text{Var}(X) = \sigma^2$ . The “standard normal” random variable is typically denoted  $Z$  and has mean 0 and variance 1: if  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ . The CDF has no closed form, but we denote the CDF of the standard normal as  $\Phi(z) = F_Z(z) = \mathbb{P}(Z \leq z)$ . Note from symmetry of the probability density function about  $z = 0$  that:  $\Phi(-z) = 1 - \Phi(z)$ .

**Closure of the Normal Under Scale and Shift:** If  $X \sim \mathcal{N}(\mu, \sigma^2)$ , then  $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$ . In particular, we can always scale/shift to get the standard Normal:  $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ .

**Closure of the Normal Under Addition:** If  $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$  and  $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$  are independent, then

$$aX + bY + c \sim \mathcal{N}(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

**4.4 Transforming Continuous RVs**

**Steps to compute PDF of  $Y = g(X)$  from  $X$  (via CDF):** Suppose  $X$  is a continuous RV.

1. Write down the range  $\Omega_X$ , PDF  $f_X$ , and CDF  $F_X$ .
2. Compute the range  $\Omega_Y = \{g(x) : x \in \Omega_X\}$ .
3. Start computing the CDF of  $Y$  on  $\Omega_Y$ ,  $F_Y(y) = \mathbb{P}(g(X) \leq y)$ , in terms of  $F_X$ .
4. Differentiate the CDF  $F_Y(y)$  to get the PDF  $f_Y(y)$  on  $\Omega_Y$ .  $f_Y$  is 0 outside  $\Omega_Y$ .

**Explicit Formula to compute PDF of  $Y = g(X)$  from  $X$  (Univariate Case):** Suppose  $X$  is a *continuous* RV. If  $Y = g(X)$  and  $g : \Omega_X \rightarrow \Omega_Y$  is *strictly monotone* and *invertible* with inverse  $X = g^{-1}(Y) = h(Y)$ , then

$$f_Y(y) = \begin{cases} f_X(h(y)) \cdot |h'(y)| & \text{if } y \in \Omega_Y \\ 0 & \text{otherwise} \end{cases}$$

**Explicit Formula to compute PDF of  $Y = g(X)$  from  $X$  (Multivariate Case):** Let  $\mathbf{X} = (X_1, \dots, X_n)$ ,  $\mathbf{Y} = (Y_1, \dots, Y_n)$  be continuous random vectors (each component is a continuous rv) with the same dimension  $n$  (so  $\Omega_{\mathbf{X}}, \Omega_{\mathbf{Y}} \subseteq \mathbb{R}^n$ ), and  $\mathbf{Y} = g(\mathbf{X})$  where  $g : \Omega_{\mathbf{X}} \rightarrow \Omega_{\mathbf{Y}}$  is invertible and differentiable, with differentiable inverse  $\mathbf{X} = g^{-1}(\mathbf{y}) = h(\mathbf{y})$ . Then,

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(h(\mathbf{y})) \left| \det \left( \frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right) \right|$$

where  $\left( \frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right) \in \mathbb{R}^{n \times n}$  is the Jacobian matrix of partial derivatives of  $h$ , with

$$\left( \frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right)_{ij} = \frac{\partial (h(\mathbf{y}))_i}{\partial y_j}$$

## 5 Multiple Random Variables

### 5.1 Joint Discrete Distributions

**Cartesian Product of Sets:** The **Cartesian product** of sets  $A$  and  $B$  is denoted:  $A \times B = \{(a, b) : a \in A, b \in B\}$ .

**Joint PMFs:** Let  $X, Y$  be discrete random variables. The joint PMF of  $X$  and  $Y$  is:

$$p_{X,Y}(a, b) = \mathbb{P}(X = a, Y = b)$$

The joint range is the set of pairs  $(c, d)$  that have nonzero probability:

$$\Omega_{X,Y} = \{(c, d) : p_{X,Y}(c, d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that the probabilities in the table must sum to 1:

$$\sum_{(s,t) \in \Omega_{X,Y}} p_{X,Y}(s, t) = 1$$

Further, note that if  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  is a function, then LOTUS extends to the multidimensional case:

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

**Marginal PMFs:** Let  $X, Y$  be discrete random variables. The marginal PMF of  $X$  is:  $p_X(a) = \sum_{b \in \Omega_Y} p_{X,Y}(a, b)$ .

**Independence (DRVs):** Discrete RVs  $X, Y$  are **independent**, written  $X \perp Y$ , if for all  $x \in \Omega_X$  and  $y \in \Omega_Y$ :  $p_{X,Y}(x, y) = p_X(x)p_Y(y)$ .

**Variance Adds for Independent RVs:** If  $X \perp Y$ , then:  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .

### 5.2 Joint Continuous Distributions

**Joint PDFs:** Let  $X, Y$  be continuous random variables. The joint PDF of  $X$  and  $Y$  is:

$$f_{X,Y}(a, b) \geq 0$$

The joint range is the set of pairs  $(c, d)$  that have nonzero density:

$$\Omega_{X,Y} = \{(c, d) : f_{X,Y}(c, d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that the double integral over all values must be 1:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv = 1$$

Further, note that if  $g : \mathbb{R}^2 \rightarrow \mathbb{R}$  is a function, then LOTUS extends to the multidimensional case:

$$\mathbb{E} [g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(s, t) f_{X,Y}(s, t) ds dt$$

The joint PDF must satisfy the following (similar to univariate PDFs):

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

**Marginal PDFs:** Let  $X, Y$  be continuous random variables. The marginal PDF of  $X$  is:  $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$ .

**Independence of Continuous Random Variables:** Continuous RVs  $X, Y$  are independent, written  $X \perp Y$ , if for all  $x \in \Omega_X$  and  $y \in \Omega_Y$ ,  $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ .

### 5.3 Conditional Distributions

**Conditional PMFs and PDFs:** If  $X, Y$  are discrete, the conditional PMF of  $X$  given  $Y$  is:

$$p_{X|Y}(a | b) = \mathbb{P}(X = a | Y = b) = \frac{p_{X,Y}(a, b)}{p_Y(b)} = \frac{p_{Y|X}(b | a) p_X(a)}{p_Y(b)}$$

Similarly for continuous RVs, but with  $f$ 's instead of  $p$ 's (PDFs instead of PMFs).

**Conditional Expectation:** If  $X$  is discrete (and  $Y$  is either discrete or continuous), then we define the conditional expectation of  $g(X)$  given (the event that)  $Y = y$  as:

$$\mathbb{E} [g(X) | Y = y] = \sum_{x \in \Omega_X} g(x) p_{X|Y}(x | y)$$

If  $X$  is continuous (and  $Y$  is either discrete or continuous), then

$$\mathbb{E} [g(X) | Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x | y) dx$$

Notice that these sums and integrals are **over**  $x$  (not  $y$ ), since  $\mathbb{E} [g(X) | Y = y]$  is a function of  $y$ .

**Law of Total Expectation (LTE):** Let  $X, Y$  be jointly distributed random variables.

If  $Y$  is discrete (and  $X$  is either discrete or continuous), then:

$$\mathbb{E} [g(X)] = \sum_{y \in \Omega_Y} \mathbb{E} [g(X) | Y = y] p_Y(y)$$

If  $Y$  is continuous (and  $X$  is either discrete or continuous), then

$$\mathbb{E} [g(X)] = \int_{-\infty}^{\infty} \mathbb{E} [g(X) | Y = y] f_Y(y) dy$$

Basically, for  $\mathbb{E} [g(X)]$ , we take a weighted average of  $\mathbb{E} [g(X) | Y = y]$  over all possible values of  $y$ .

**Multivariate: Discrete to Continuous:**

	Discrete	Continuous
<b>Joint Dist</b>	$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$	$f_{X,Y}(x, y) \neq \mathbb{P}(X = x, Y = y)$
<b>Joint CDF</b>	$F_{X,Y}(x, y) = \sum_{t \leq x, s \leq y} p_{X,Y}(t, s)$	$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(t, s) ds dt$
<b>Normalization</b>	$\sum_{x,y} p_{X,Y}(x, y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$
<b>Marginal Dist</b>	$p_X(x) = \sum_y p_{X,Y}(x, y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$
<b>Expectation</b>	$\mathbb{E} [g(X, Y)] = \sum_{x,y} g(x, y) p_{X,Y}(x, y)$	$\mathbb{E} [g(X, Y)] = \int \int g(x, y) f_{X,Y}(x, y) dx dy$
<b>Conditional Dist</b>	$p_{X Y}(x y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X Y}(x y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
<b>Conditional Exp</b>	$\mathbb{E} [X Y = y] = \sum_x x p_{X Y}(x y)$	$\mathbb{E} [X Y = y] = \int_{-\infty}^{\infty} x f_{X Y}(x y) dx$
<b>Independence</b>	$\forall x, y, p_{X,Y}(x, y) = p_X(x) p_Y(y)$	$\forall x, y, f_{X,Y}(x, y) = f_X(x) f_Y(y)$

### 5.4 Covariance and Correlation

**Covariance:** The **covariance** of  $X$  and  $Y$  is:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

Covariance satisfies the following properties:

1. If  $X \perp Y$ , then  $\text{Cov}(X, Y) = 0$  (but not necessarily vice versa).
2.  $\text{Cov}(X, X) = \text{Var}(X)$ . (Just plug in  $Y = X$ ).
3.  $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ . (Multiplication is commutative).
4.  $\text{Cov}(X + c, Y) = \text{Cov}(X, Y)$ . (Shifting doesn't and shouldn't affect the covariance).
5.  $\text{Cov}(aX + bY, Z) = a \cdot \text{Cov}(X, Z) + b \cdot \text{Cov}(Y, Z)$ . This can be easily remembered like the distributive property of scalars  $(aX + bY)Z = a(XZ) + b(YZ)$ .
6.  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$ , and hence if  $X \perp Y$ , then  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ .
7.  $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$ . That is covariance works like FOIL (first, outer, inner, last) for multiplication of sums  $((a + b + c)(d + e) = ad + ae + bd + be + cd + ce)$ .

**(Pearson) Correlation:** The **(Pearson) correlation** of  $X$  and  $Y$  is:  $\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$ .

It is always true that  $-1 \leq \rho(X, Y) \leq 1$ . That is, correlation is just a normalized version of covariance. Most notably,  $\rho(X, Y) = \pm 1$  if and only if  $Y = aX + b$  for some constants  $a, b \in \mathbb{R}$ , and then the sign of  $\rho$  is the same as that of  $a$ .

**Variance of Sums of RVs:** Let  $X_1, \dots, X_n$  be any RVs (independent or not). Then,

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$$

### 5.5 Convolution

**Law of Total Probability for Random Variables:**

Discrete version: If  $X, Y$  are discrete:

$$p_X(x) = \sum_y p_{X,Y}(x, y) = \sum_y p_{X|Y}(x | y)p_Y(y)$$

Continuous version: If  $X, Y$  are continuous:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x | y)f_Y(y)dy$$

**Convolution:** Let  $X, Y$  be *independent* RVs, and  $Z = X + Y$ .

Discrete version: If  $X, Y$  are discrete:

$$p_Z(z) = \sum_{x \in \Omega_X} p_X(x)p_Y(z - x)$$

Continuous version: If  $X, Y$  are continuous:

$$f_Z(z) = \int_{x \in \Omega_X} f_X(x)f_Y(z - x)dx$$

### 5.6 Limit Theorems

**The Sample Mean + Properties:** Let  $X_1, X_2, \dots, X_n$  be a sequence of iid RVs with mean  $\mu$  and variance  $\sigma^2$ . The **sample mean** is:  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ . Further,  $\mathbb{E}[\bar{X}_n] = \mu$  and  $\text{Var}(\bar{X}_n) = \sigma^2/n$

**The Law of Large Numbers (LLN):** Let  $X_1, \dots, X_n$  be iid RVs with the same mean  $\mu$ . As  $n \rightarrow \infty$ , the sample mean



$\bar{X}_n$  converges to the true mean  $\mu$ .

**The Central Limit Theorem (CLT):** Let  $X_1, \dots, X_n$  be a sequence of iid RVs with mean  $\mu$  and (finite) variance  $\sigma^2$ . Then as  $n \rightarrow \infty$ ,

$$\bar{X}_n \rightarrow \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The mean or variance are not a surprise; the importance of the CLT is, regardless of the distribution of  $X_i$ 's, the sample mean approaches a Normal distribution as  $n \rightarrow \infty$ .

**The Continuity Correction:** When approximating an integer-valued (*discrete*) random variable  $X$  with a *continuous* one  $Y$  (such as in the CLT), if asked to find a  $\mathbb{P}(a \leq X \leq b)$  for integers  $a \leq b$ , you should use  $\mathbb{P}(a - 0.5 \leq Y \leq b + 0.5)$  so that the width of the interval being integrated is the same as the number of terms summed over ( $b - a + 1$ ).

## 6 Concentration Inequalities

### 6.1 Markov and Chebyshev Inequalities

**Markov's Inequality:** Let  $X \geq 0$  be a **non-negative** RV, and let  $k > 0$ . Then:  $\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$ .

**Chebyshev's Inequality:** Let  $X$  be any RV with expected value  $\mu = \mathbb{E}[X]$  and finite variance  $\text{Var}(X)$ . Then, for any real number  $\alpha > 0$ . Then,  $\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\text{Var}(X)}{\alpha^2}$ .

### 6.2 The Chernoff Bound

**(Multiplicative) Chernoff Bound:**

Let  $X_1, X_2, \dots, X_n$  be *independent* Bernoulli random variables.

Let  $X = \sum_{i=1}^n X_i$ , and  $\mu = \mathbb{E}[X]$ . Then, for any  $0 \leq \delta \leq 1$ ,

$$\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{3}\right) \quad \text{and} \quad \mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\left(-\frac{\delta^2 \mu}{2}\right)$$

### 6.3 Even More Inequalities

**The Union Bound:** Let  $E_1, E_2, \dots, E_n$  be a collection of events. Then:  $\mathbb{P}(\bigcup_{i=1}^n E_i) \leq \sum_{i=1}^n \mathbb{P}(E_i)$ .

A similar statement also holds if the number of events is *countably* infinite.

**Hoeffding's Inequality:** Let  $X_1, \dots, X_n$  be independent random variables, where each  $X_i$  is bounded:  $a_i \leq X_i \leq b_i$  and let  $\bar{X}_n$  be their sample mean. Then,

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}[\bar{X}_n]| \geq t) \leq 2 \exp\left(-\frac{2n^2 t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

In the case  $X_1, \dots, X_n$  are iid (so  $a \leq X_i \leq b$  for all  $i$ ) with mean  $\mu$ , then

$$\mathbb{P}(|\bar{X}_n - \mu| \geq t) \leq 2 \exp\left(-\frac{2n^2 t^2}{n(b-a)^2}\right) = 2 \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

## 7 Statistical Estimation

### 7.1 Maximum Likelihood Estimation

**Realization / Sample:** A **realization/sample**  $x$  of a random variable  $X$  is the value that is actually observed (will always be in  $\Omega_X$ ).

**Likelihood:** Let  $\mathbf{x} = (x_1, \dots, x_n)$  be iid realizations from PMF  $p_X(t | \theta)$  (if  $X$  is discrete), or from density  $f_X(t | \theta)$  (if  $X$  is continuous), where  $\theta$  is a parameter (or vector of parameters). We define the **likelihood** of  $\mathbf{x}$  given  $\theta$  to be the "probability" of observing  $\mathbf{x}$  if the true parameter is  $\theta$ . The **log-likelihood** is just the log of the likelihood, which is typically easier to optimize.

If  $X$  is discrete,

$$L(\mathbf{x} | \theta) = \prod_{i=1}^n p_X(x_i | \theta) \quad \ln L(\mathbf{x} | \theta) = \sum_{i=1}^n \ln p_X(x_i | \theta)$$

If  $X$  is continuous,

$$L(\mathbf{x} | \theta) = \prod_{i=1}^n f_X(x_i | \theta) \quad \ln L(\mathbf{x} | \theta) = \sum_{i=1}^n \ln f_X(x_i | \theta)$$

**Maximum Likelihood Estimator (MLE):** Let  $\mathbf{x} = (x_1, \dots, x_n)$  be iid realizations from probability mass function  $p_X(t | \theta)$  (if  $X$  is discrete), or from density  $f_X(t | \theta)$  (if  $X$  is continuous), where  $\theta$  is a parameter (or vector of parameters). We define the **maximum likelihood estimator (MLE)**  $\hat{\theta}_{MLE}$  of  $\theta$  to be the parameter which maximizes the likelihood/log-likelihood:

$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(\mathbf{x} | \theta) = \arg \max_{\theta} \ln L(\mathbf{x} | \theta)$$

## 7.2 Properties of Estimators

**Bias:** Let  $\hat{\theta}$  be an estimator for  $\theta$ . The **bias** of  $\hat{\theta}$  as an estimator for  $\theta$  is  $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}[\hat{\theta}] - \theta$ . If  $\text{Bias}(\hat{\theta}, \theta) = 0$ , or equivalently  $\mathbb{E}[\hat{\theta}] = \theta$ , then we say  $\hat{\theta}$  is an **unbiased** estimator of  $\theta$ .

# 8 Statistical Inference

## 8.1 Confidence Intervals

**Confidence Interval:** Suppose you have iid samples  $x_1, \dots, x_n$  from some distribution with unknown parameter  $\theta$ , and you have some estimator  $\hat{\theta}$  for  $\theta$ .

A  $100(1 - \alpha)\%$  **confidence interval** for  $\theta$  is an interval (typically but not always) centered at  $\hat{\theta}$ ,  $[\hat{\theta} - \Delta, \hat{\theta} + \Delta]$ , such that the probability (over the randomness in the samples  $x_1, \dots, x_n$ )  $\theta$  lies in the interval is  $1 - \alpha$ :

$$\mathbb{P}(\theta \in [\hat{\theta} - \Delta, \hat{\theta} + \Delta]) = 1 - \alpha$$

If  $\hat{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$  is the sample mean, then  $\hat{\theta}$  is approximately normal by the CLT, and a  $100(1 - \alpha)\%$  confidence interval is given by the formula:

$$\left[ \hat{\theta} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \hat{\theta} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

where  $z_{1-\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$  and  $\sigma$  is the true standard deviation of a single sample (which may need to be estimated).