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, $\ldots$, and N_m outcomes for the m^{th} stage, then the total number of outcomes of the experiment is $N_1 \times N_2 \cdot \cdots \cdot N_m = \prod_{i=1}^{m} N_i$.

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

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

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:

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

***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

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

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 Ω 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)** $\mathbb{P}\left(E^C\right) = 1 - \mathbb{P}(E)$

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

2.2 Conditional Probability

Conditional Probability: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Bayes Theorem: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$

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

- **(Exhaustive)** $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^n E_i = \Omega$ (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 Ω.

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

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

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

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

2.3 Independence

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

$$\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 A_2) \cdots \mathbb{P}(A_n \mid A_1, \ldots, A_{n-1})$$

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

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

Mutual Independence: We say n events $A_1, A_2, \ldots, A_n$ are **(mutually) independent** if, for *any* subset $I \subseteq [n] = \{1, 2, \ldots, n\}$, 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. $\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$

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 $X : \Omega \to \mathbb{R}$. The set of possible values X can take on is its **range/support**, denoted Ω_X.

If Ω_X is finite or countable infinite (typically integers or a subset), X is a **discrete RV**. Else if Ω_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 $p_X : \Omega_X \to [0,1]$ 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 $X_1, \ldots, X_n$, 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: $\mathsf{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

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

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

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 $X_1, \ldots, X_n$ 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 $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.

X **Bernoulli Process:** A **Bernoulli process** with parameter p is a sequence of independent coin flips $X_1, X_2, X_3, \ldots$ 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 $\mathsf{Var}(X) = p(1-p)$. 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: $X \sim \text{Binomial}(n, p)$ ($\mathsf{Bin}(n, p)$ for short) iff X has PMF

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

$\mathbb{E}[X] = np$ and $\mathsf{Var}(X) = np(1-p)$. 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 $n \to \infty$ and $p \to$

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

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 $a \le b$, iff X has PMF:

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

$\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathsf{Var}(X) = \frac{(b-a)(b-a+2)}{12}$. 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:

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

$\mathbb{E}[X] = \frac{1}{p}$ and $\mathsf{Var}(X) = \frac{1-p}{p^2}$. 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:

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

$\mathbb{E}[X] = \frac{r}{p}$ and $\mathsf{Var}(X) = \frac{r(1-p)}{p^2}$. 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 $X_1, \ldots, X_n$ are independent Negative Binomial RV's, where $X_i \sim \text{NegBin}(r_i, p)$, then $X = X_1 + \ldots + X_n \sim \text{NegBin}(r_1 + \ldots + r_n, p)$.

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:

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

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

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

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

$\mathbb{E}[X] = n\frac{K}{N}$ and $\mathsf{Var}(X) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$. 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 $\text{Bin}\left(n, \frac{K}{N}\right)$.

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 $f_X : \mathbb{R} \to \mathbb{R}$, such that the following properties hold:

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

Cumulative Distribution Function (CDF): The **cumulative distribution function (CDF)** of ANY random variable (discrete or continuous) is defined to be the function $F_X : \mathbb{R} \to \mathbb{R}$ with $F_X(t) = \mathbb{P}(X \le t)$. 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): $X \sim \text{Uniform}(a, b)$ ($\text{Unif}(a, b)$ 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: $X \sim \text{Exponential}(\lambda)$ ($\text{Exp}(\lambda)$ 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:

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

X

Gamma Random Variable: $X \sim \text{Gamma}(r, \lambda)$ ($\text{Gam}(r, \lambda)$ for short) iff X has PDF:

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x \in \Omega_X = [0, \infty)$$

$\mathbb{E}[X] = \frac{r}{\lambda}$ and $\text{Var}(X) = \frac{r}{\lambda^2}$. X is the sum of r iid $\text{Exp}(\lambda)$ random variables. In the above PDF, for positive integers r, $\Gamma(r) = (r-1)!$ (a normalizing constant). An example of a Gamma RV is the waiting time until the r-th event in the Poisson process. If $X_1, \ldots, X_n$ are independent Gamma RV's, where $X_i \sim \text{Gam}(r_i, \lambda)$, then $X = X_1 + \ldots + X_n \sim \text{Gam}(r_1 + \ldots + r_n, \lambda)$. It also serves as a conjugate prior for λ in the Poisson and Exponential distributions.

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 Ω_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 Ω_Y, $F_Y(y) = \mathbb{P}(g(X) \le y)$, in terms of F_X.
4. Differentiate the CDF $F_Y(y)$ to get the PDF $f_Y(y)$ on Ω_Y. f_Y is 0 outside Ω_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 $g : \Omega_X \to \Omega_Y$ 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 $\mathbf{X} = (X_1, ..., X_n)$, $\mathbf{Y} = (Y_1, ..., Y_n)$ 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 $g : \Omega_{\mathbf{X}} \to \Omega_{\mathbf{Y}}$ 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

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

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 $g : \mathbb{R}^2 \to \mathbb{R}$ 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: $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.

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

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

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

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

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

sort of

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

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

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:

sort of

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

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

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

Notice that these sums and integrals are **over** x (not y), since $\mathbb{E}\left[g(X) \mid Y = y\right]$ 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:

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

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

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

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

Multivariate: Discrete to Continuous:

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

5.4 Covariance and Correlation

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

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

Covariance satisfies the following properties:

1. If $X \perp Y$, then $\mathsf{Cov}(X,Y) = 0$ (but not necessarily vice versa).
2. $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$. (Just plug in $Y = X$).
3. $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$. (Multiplication is commutative).
4. $\mathsf{Cov}(X + c, Y) = \mathsf{Cov}(X,Y)$. (Shifting doesn't and shouldn't affect the covariance).
5. $\mathsf{Cov}(aX + bY, Z) = a \cdot \mathsf{Cov}(X)\, Z + b \cdot \mathsf{Cov}(Y,Z)$. This can be easily remembered like the distributive property of scalars $(aX + bY)Z = a(XZ) + b(YZ)$.
6. $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$, and hence if $X \perp Y$, then $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.
7. $\mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathsf{Cov}(X_i, Y_j)$. 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: $\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}}$.
It is always true that $-1 \le \rho(X,Y) \le 1$. 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 ρ is the same as that of a.

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

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

5.5 Convolution

Law of Total Probability for Random Variables:
Discrete version: If X, Y are discrete:

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

Continuous version: If X, Y are continuous:

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

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 Moment Generating Functions

Moments: Let X be a random variable and $c \in \mathbb{R}$ a scalar. Then: The **k-th moment** of X is $\mathbb{E}\left[X^k\right]$ and the k-th moment of X (about c) is: $\mathbb{E}\left[(X-c)^k\right]$.

Moment Generating Functions (MGFs): The **moment generating function (MGF)** of X is a function of a dummy variable t (use LOTUS to compute this): $M_X(t) = \mathbb{E}\left[e^{tX}\right]$.

Properties and Uniqueness of Moment Generating Functions: For a function $f : \mathbb{R} \to \mathbb{R}$, we will denote $f^{(n)}(x)$ to be the n-th derivative of $f(x)$. Let X, Y be *independent* random variables, and $a, b \in \mathbb{R}$ be scalars. Then MGFs satisfy the following properties:

1. $M_X'(0) = \mathbb{E}[X]$, $M_X''(0) = \mathbb{E}\left[X^2\right]$, and in general $M_X^{(n)} = \mathbb{E}[X^n]$. This is why we call M_X a *moment generating* function, as we can use it to generate the moments of X.
2. $M_{aX+b}(t) = e^{tb}M_X(at)$.
3. If $X \perp Y$, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.
4. (**Uniqueness**) The following are equivalent:
   (a) X and Y have the same distribution.
   (b) $f_X(z) = f_Y(z)$ for all $z \in \mathbb{R}$.
   (c) $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$.
   (d) There is an $\varepsilon > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$.

   That is M_X uniquely identifies a distribution, just like PDFs/PMFs or CDFs do.

5.7 Limit Theorems

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

The Law of Large Numbers (LLN): Let $X_1, \ldots, X_n$ be iid RVs with the same mean μ. As $n \to \infty$, the sample mean $\overline{X}_n$ converges to the true mean μ.

The Central Limit Theorem (CLT): Let $X_1, \ldots X_n$ be a sequence of iid RVs with mean μ and (finite) variance σ^2. Then as $n \to \infty$,

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

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 $n \to \infty$.

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 $(b - a + 1)$.

5.8 The Multinomial Distribution

Random Vectors (RVTRs): Let $X_1, ..., X_n$ be random variables. We say $\mathbf{X} = (X_1, \ldots, X_n)^T$ is a **random vector**. Expectation is defined pointwise: $\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^T$.

Covariance Matrices: The **covariance matrix** of a random vector $\mathbf{X} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ is the matrix $\Sigma = \mathsf{Var}(\mathbf{X}) =$

$\mathsf{Cov}(\mathbf{X})$ whose entries $\Sigma_{ij} = \mathsf{Cov}(X_i, X_j)$. The formula for this is:

$$\Sigma = \mathsf{Var}(\mathbf{X}) = \mathsf{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathbf{T}}\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^{\mathbf{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathbf{T}}$$
$$= \begin{bmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1, X_2) & \dots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathsf{Var}(X_2) & \dots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \dots & \mathsf{Var}(X_n) \end{bmatrix}$$

Notice that the covariance matrix is **symmetric** ($\Sigma_{ij} = \Sigma_{ji}$), and has variances on the diagonal.

The Multinomial Distribution: Suppose there are r outcomes, with probabilities $\mathbf{p} = (p_1, p_2, ..., p_r)$ respectively, such that $\sum_{i=1}^{r} p_i = 1$. Suppose we have n independent trials, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$ be the rvtr of counts of each outcome. Then, we say $\mathbf{Y} \sim \text{Mult}_r(n, \mathbf{p})$:
The joint PMF of $\mathbf{Y}$ is:

$$p_{Y_1,...,Y_r}(k_1, ...k_r) = \binom{n}{k_1, ..., k_r} \prod_{i=1}^{r} p_i^{k_i}, \quad k_1, ...k_r \geq 0 \text{ and } \sum_{i=1}^{r} k_i = n$$

Notice that each Y_i is marginally $\text{Bin}(n, p_i)$. Hence, $\mathbb{E}[Y_i] = np_i$ and $\mathsf{Var}(Y_i) = np_i(1-p_i)$.
Then, we can specify the entire mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix:

$$\mathbb{E}[\mathbf{Y}] = n\mathbf{p} = \begin{bmatrix} np_1 \\ \vdots \\ np_r \end{bmatrix} \qquad \mathsf{Var}(Y_i) = np_i(1-p_i) \qquad \mathsf{Cov}(Y_i, Y_j) = -np_ip_j$$

The Multivariate Hypergeometric (MVHG) Distribution: Suppose there are r different colors of balls in a bag, having $\mathbf{K} = (K_1, ..., K_r)$ balls of each color, $1 \leq i \leq r$. Let $N = \sum_{i=1}^{r} K_i$ be the total number of balls in the bag, and suppose we draw n without replacement. Let $\mathbf{Y} = (Y_1, ..., Y_r)$ be the rvtr such that Y_i is the number of balls of color i we drew. We write that $\mathbf{Y} \sim \text{MVHG}_r(N, \mathbf{K}, n)$ The joint PMF of Y is:

$$p_{Y_1,...,Y_r}(k_1, ...k_r) = \frac{\prod_{i=1}^{r} \binom{K_i}{k_i}}{\binom{N}{n}}, \quad 0 \leq k_i \leq K_i \text{ for all } 1 \leq i \leq r \text{ and } \sum_{i=1}^{r} k_r = n$$

Notice that each Y_i is marginally $\text{HypGeo}(N, K_i, n)$, so $\mathbb{E}[Y_i] = n\frac{K_i}{N}$ and
$\mathsf{Var}(Y_i) = n\dfrac{K_i}{N} \cdot \dfrac{N-K_i}{N} \cdot \dfrac{N-n}{N-1}$. The mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix are:

$$\mathbb{E}[\mathbf{Y}] = n\frac{\mathbf{K}}{N} = \begin{bmatrix} n\frac{K_1}{N} \\ \vdots \\ n\frac{K_r}{N} \end{bmatrix} \qquad \mathsf{Var}(Y_i) = n\frac{K_i}{N} \cdot \frac{N-K_i}{N} \cdot \frac{N-n}{N-1} \qquad \mathsf{Cov}(Y_i, Y_j) = -n\frac{K_i}{N}\frac{K_j}{N} \cdot \frac{N-n}{N-1}$$

5.9 The Multivariate Normal Distribution

Properties of Expectation and Variance Hold for RVTRs: Let $\mathbf{X}$ be an n-dimensional RVTR, $A \in \mathbb{R}^{n \times n}$ be a constant matrix, $\mathbf{b} \in \mathbb{R}^n$ be a constant vector. Then: $\mathbb{E}[A\mathbf{X} + \mathbf{b}] = A\mathbb{E}[\mathbf{X}] + \mathbf{b}$ and $\mathsf{Var}(A\mathbf{X} + \mathbf{b}) = A\mathsf{Var}(\mathbf{X})A^T$.

The Multivariate Normal Distribution: A random vector $\mathbf{X} = (X_1, ..., X_n)$ has a multivariate Normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and (symmetric and positive-definite) covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, written $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, if it has the following joint PDF:

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right), \qquad \boldsymbol{x} \in \mathbb{R}^n$$

Additionally, let us recall that for any RVs X and Y: $X \perp Y \to \mathsf{Cov}(X, Y) = 0$. If $\mathbf{X} = (X_1, \ldots, X_n)$ is Multivariate Normal, the converse also holds: $\mathsf{Cov}(X_i, X_j) = 0 \to X_i \perp X_j$.

5.10 Order Statistics

Order Statistics: Suppose Y_1, ..., Y_n are iid *continuous* random variables with common PDF f_Y and common CDF F_Y. We sort the Y_i's such that $Y_{\min} \equiv Y_{(1)} < Y_{(2)} < ... < Y_{(n)} \equiv Y_{\max}$.

Notice that we can't have equality because with continuous random variables, the probability that any two are equal is 0. Notice that each $Y_{(i)}$ is a random variable as well! We call $Y_{(i)}$ the **ith order statistic**, i.e. the ith smallest in a sample of size n. The density function of each $Y_{(i)}$ is

$$f_{Y_{(i)}}(y) = \binom{n}{i-1, 1, n-i} \cdot [F_Y(y)]^{i-1} \cdot [1 - F_Y(y)]^{n-i} \cdot f_Y(y), y \in \Omega_Y$$

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: $\mathbb{P}(X \geq k) \leq \dfrac{\mathbb{E}[X]}{k}$.

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

6.2 The Chernoff Bound

Chernoff Bound for Binomial: Let $X \sim \text{Bin}(n, p)$ and let $\mu = \mathbb{E}[X]$. For any $0 < \delta < 1$:

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

6.3 Even More Inequalities

The Union Bound: Let $E_1, E_2, ..., E_n$ be a collection of events. Then: $\mathbb{P}\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n \mathbb{P}(E_i)$.
A similar statement also holds if the number of events is *countably* infinite.

Convex Sets: A set $S \subseteq \mathbb{R}^n$ is a **convex set** if for any $x_1, \ldots, x_m \in S$

$$\left\{\sum_{i=1}^m p_i x_i : p_1, ..., p_m \geq 0 \text{ and } \sum_{i=1}^m p_i = 1\right\} \subseteq S$$

Convex Functions: Let $S \subseteq \mathbb{R}^n$ be a convex set. A function $g : S \to \mathbb{R}$ is a **convex function** if for any $x_1, ..., x_m \in S$, and $p_1, ..., p_m \geq 0$ such that $\sum_{i=1}^m p_i = 1$,

$$g\left(\sum_{i=1}^m p_i x_i\right) \leq \sum_{i=1}^m p_i g(x_i)$$

Jensen's Inequality: Let X be any RV, and $g : \mathbb{R} \to \mathbb{R}$ be convex. Then, $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$.

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

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

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

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

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 Ω_X).

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

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

If X is continuous,

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

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

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

7.2 MLE Examples

7.3 Method of Moments Estimation

Sample Moments: Let X be a random variable, and $c \in \mathbb{R}$ a scalar. Let $x_1, \dots, x_n$ be iid realizations (samples) from X.
The k^{th} **sample moment of** X is: $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.
The k^{th} **sample moment of** X **(about** c**)** is: $\frac{1}{n}\sum_{i=1}^{n} (x_i - c)^k$.

Method of Moments Estimation: Let $x = (x_1, \dots, x_n)$ be iid realizations (samples) from PMF $p_X(t; \theta)$ (if X is discrete), or from density $f_X(t; \theta)$ (if X continuous), where θ is a parameter (or vector of parameters).

We then define the **Method of Moments (MoM)** estimator $\hat{\theta}_{MoM}$ of $\theta = (\theta_1, \dots, \theta_k)$ to be a solution (if it exists) to the k simultaneous equations where, for $j = 1, \dots, k$, we set the j^{th} true and sample moments equal:

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \cdots \qquad \mathbb{E}\left[X^k\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

7.4 The Beta and Dirichlet Distributions

Beta Random Variable: $X \sim \text{Beta}(\alpha, \beta)$, if and only if X has the following PDF:

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & x \in \Omega_X = [0,1] \\ 0, & \text{otherwise} \end{cases}$$

X is typically the belief distribution about some unknown probability of success, where we pretend we've seen $\alpha - 1$ successes and $\beta - 1$ failures. Hence the mode (most likely value of the probability/point with highest density) $\arg\max_{x \in [0,1]} f_X(x)$, is

$$\text{mode}[X] = \frac{\alpha-1}{(\alpha-1)+(\beta-1)}$$

Also note that there is an annoying "off-by-1" issue: ($\alpha - 1$ heads and $\beta - 1$ tails), so when choosing these parameters, be careful! It also serves as a conjugate prior for p in the Bernoulli and Geometric distributions.

Dirichlet RV: $X \sim \text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_r)$, if and only if X has the following density function:

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{r} x_i^{a_i-1}, & x_i \in (0,1) \text{ and } \sum_{i=1}^{r} x_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

This is a generalization of the Beta random variable from 2 outcomes to r. The random vector X is typically the belief distribution about some unknown probabilities of the different outcomes, where we pretend we saw $a_1 - 1$ outcomes of type 1, $a_2 - 1$ outcomes of type 2, ..., and $a_r - 1$ outcomes of type r . Hence, the mode of the distribution is the vector, $\arg\max_{x \in [0,1]^d \text{ and } \sum x_i = 1} f_{\mathbf{X}}(\mathbf{x})$, is

$$\text{mode}[\mathbf{X}] = \left(\frac{\alpha_1 - 1}{\sum_{i=1}^{r}(a_i - 1)}, \frac{\alpha_2 - 1}{\sum_{i=1}^{r}(a_i - 1)}, \ldots, \frac{\alpha_r - 1}{\sum_{i=1}^{r}(a_i - 1)}\right)$$

7.5 Maximum A Posteriori Estimation

Maximum A Posteriori (MAP) Estimation: Let $x = (x_1, \ldots, x_n)$ be iid realizations from PMF $p_X(t\,; \Theta = \theta)$ (if X discrete), or from density $f_X(t\,; \Theta = \theta)$ (if X continuous), where Θ is the random variable representing the parameter (or vector of parameters). We define the **Maximum A Posteriori (MAP)** estimator $\hat{\theta}_{MAP}$ of Θ to be the parameter which maximizes the **posterior** distribution of Θ given the data (the mode).

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ \pi_\Theta(\theta \mid \mathbf{x}) = \underset{\theta}{\text{argmax}}\ L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$$

7.6 Properties of Estimators I

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

X

Mean Squared Error (MSE): The **mean squared error (MSE)** of an estimator $\hat{\theta}$ of θ is $\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$. If $\hat{\theta}$ is an unbiased estimator of θ (i.e. $\mathbb{E}\left[\hat{\theta}\right] = \theta$), then you can see that $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right)$. In fact, in general $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2$.

7.7 Properties of Estimators II

Consistency: An estimator $\hat{\theta}_n$ (depending on n iid samples) of θ is said to be **consistent** if it converges (in probability) to θ. That is, for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$.

X

Fisher Information: Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from PMF $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). The **Fisher Information** of a parameter θ is defined to be

$$I(\theta) = n \cdot \mathbb{E}\left[\left(\frac{\partial \ln L(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]$$

X

Cramer-Rao Lower Bound (CRLB): Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from PMF $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). If $\hat{\theta}$ is an *unbiased* estimator for θ, then

$$\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) \geq \frac{1}{I(\theta)}$$

That is, for any unbiased estimator $\hat{\theta}$ for θ, the variance (=MSE) is at least $\frac{1}{I(\theta)}$. If we achieve this lower bound, meaning our variance is exactly equal to $\frac{1}{I(\theta)}$, then we have the best variance possible for our estimate. Hence, it is the **minimum variance unbiased estimator (MVUE)** for θ.

X

Efficiency: Let $\hat{\theta}$ be an unbiased estimator of θ. The efficiency of $\hat{\theta}$ is $e(\hat{\theta}, \theta) = \dfrac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} \leq 1$.

An estimator is said to be **efficient** if it achieves the CRLB - meaning $e(\hat{\theta}, \theta) = 1$.

X

7.8 Properties of Estimators III

Statistic: A **statistic** is any function $T : \mathbb{R}^n \to \mathbb{R}$ of samples $\mathbf{x} = (x_1, \ldots, x_n)$. For example, $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$ (the sum), $T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ (the max/largest value), $T(x_1, \ldots, x_n) = x_1$ (just take the first sample)

Sufficiency: A statistic $T = T(X_1, \ldots, X_n)$ is a **sufficient statistic** if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ and θ does not depend on θ.

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta\right) = \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \mid T = t\right)$$

Neyman-Fisher Factorization Criterion (NFFC): Let $x_1, \ldots, x_n$ be iid random samples with likelihood $L(x_1, \ldots, x_n|\theta)$. A statistic $T = T(x_1, \ldots, x_n)$ is sufficient if and only if there exist non-negative functions g and h such that:

$$L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(T(x_1, \ldots, x_n), \theta)$$

8 Statistical Inference

8.1 Confidence Intervals

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

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

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

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

8.2 Credible Intervals

X

Credible Intervals: Suppose you have iid samples $\mathbf{x} = (x_1, ..., x_n)$ from some distribution with unknown parameter Θ. You are in the **Bayesian setting**, so you have chosen a prior distribution for the RV Θ.

A $100(1-\alpha)\%$ **credible interval** for Θ is an interval $[a, b]$ such that the probability (over the randomness in Θ) that Θ lies in the interval is $1 - \alpha$:

$$P(\Theta \in [a, b]) = 1 - \alpha$$

If we've chosen the appropriate conjugate prior for the sampling distribution (like Beta for Bernoulli), the posterior is easy to compute. Say the CDF of the posterior is F_Y. Then, a $100(1-\alpha)\%$ credible interval is given by

$$\left[F_Y^{-1}\left(\frac{\alpha}{2}\right), F_Y^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

X

8.3 Introduction to Hypothesis Testing

Hypothesis Testing Procedure:

1. Make a claim (like "Airplane food is good", "Pineapples belong on pizza", etc...)
2. Set up a null hypothesis H_0 and alternative hypothesis H_A.
   (a) Alternative hypothesis can be one-sided or two-sided.
   (b) The null hypothesis is usually a "baseline", "no effect", or "benefit of the doubt".
   (c) The alternative is what you want to "prove", and is opposite the null.
3. Choose a significance level α (usually $\alpha = 0.05$ or 0.01).
4. Collect data.
5. Compute a p-value, $p = \mathbb{P}(\text{observing data at least as extreme as ours} \mid H_0 \text{ is true})$.
6. State your conclusion. Include an interpretation in the context of the problem.
   (a) If $p < \alpha$, "reject" the null hypothesis H_0 in favor of the alternative H_A.
   (b) Otherwise, "fail to reject" the null hypothesis H_0.

This quarter – we made it!

- You learned foundations of probability & some stats

  What next?

  421 algorithms
  422 modern algorithmic toolbox
  426 crypto
  431 complexity
  446 ML
  447 NLP
  493Q Quantum computation

- You learned about some applications
  - Naive Bayes spam filtering
  - Bloom filters
  - Distinct elts & Minhash
  - Markov chains & Pagerank
  - Auctions

- You learned a bit of Python

  great headstart for 446

- Many of you learned Latex

- You asked a lot of questions

Special shoutout to our wonderful TAs!!

Thanks for a great quarter!

Please fill out evaluations now!