CS 109: Probability for Computer Scientists

Section 4: Definitions and Theorems

0. Counting and Probability

(a) **Product Rule:** Suppose there are m_1 possible outcomes for event A_1, then m_2 possible outcomes for event A_2,..., m_n possible outcomes for event A_n. Then there are $m_1 \cdot m_2 \cdot m_3 \cdots m_n = \prod_{i=1}^{n} m_i$ possible outcomes overall.

(b) **Number of ways to order n distinct objects:** $n! = n \cdot (n-1) \cdots 3 \cdot 2 \cdot 1$

(c) **Number of ways to select from n distinct objects:**

   (a) **Permutations** (number of ways to linearly arrange k objects out of n distinct objects, when the order of the k objects matters):

$$\frac{n!}{(n-k)!}$$

   (b) **Combinations** (number of ways to choose k objects out of n distinct objects, when the order of the k objects does not matter):

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

(d) **Multinomial coefficients:** Suppose there are n objects, but only k are distinct, with $k \leq n$. (For example, "godoggy" has $n = 7$ objects (characters) but only $k = 4$ are distinct: (g, o, d, y)). Let n_i be the number of times object i appears, for $i \in \{1, 2, \ldots, k\}$. (For example, $(3, 2, 1, 1)$, continuing the "godoggy" example.) The number of distinct ways to arrange the n objects is:

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

(e) **Binomial Theorem:** $\forall x, y \in \mathbb{R}, \forall n \in \mathbb{N}$: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$

(f) **Principle of Inclusion-Exclusion (PIE):** 2 events: $|A \cup B| = |A| + |B| - |A \cap B|$
3 events: $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$
In general: +singles - doubles + triples - quads + ...

(g) **Pigeonhole Principle:** If there are n pigeons with k holes and $n > k$, then at least one hole contains at least 2 (or to be precise, $\lceil \frac{n}{k} \rceil$) pigeons.

(h) **Complementary Counting (Complementing):** If asked to find the number of ways to do X, you can: find the total number of ways and then subtract the number of ways to not do X.

(i) **Key Probability Definitions**

   (a) **Sample Space:** The set of all possible outcomes of an experiment, denoted S
   (b) **Event:** Some subset of the sample space, usually a capital letter such as $E \subseteq S$
   (c) **Union:** The union of two events E and F is denoted $E \cup F$
   (d) **Intersection:** The intersection of two events E and F is denoted $E \cap F$ or EF
   (e) **Mutually Exclusive:** Events E and F are mutually exclusive iff $E \cap F = \emptyset$
   (f) **Complement:** The complement of an event E is denoted E^C or $\overline{E}$, and is equal to $S \setminus E$

(g) **DeMorgan's Laws:** $(E \cup F)^C = E^C \cap F^C$ and $(E \cap F)^C = E^C \cup F^C$

(h) **Probability of an event E:** denoted $P(E)$

(i) **Partition:** Nonempty events $E_1, \ldots, E_n$ partition the sample space S iff
  - $E_1, \ldots, E_n$ are exhaustive: $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^{n} E_i = S$, and
  - $E_1, \ldots, E_n$ are pairwise mutually exclusive: $\forall i \neq j, E_i \cap E_j = \emptyset$
    - Note that for any event A (with $A \neq \emptyset, A \neq S$): A and A^C partition S

(j) **Axioms of Probability and their Consequences**

  (a) **Axiom 1: Non-negativity** For any event E, $P(E) \geq 0$

  (b) **Axiom 2: Normalization** $P(S) = 1$

  (c) **Axiom 3: Countable Additivity** If E and F are mutually exclusive, then $P(E \cup F) = P(E) + P(F)$. Also, if $E_1, E_2, ...$ is a countable sequence of disjoint events, $P(\bigcup_{k=1}^{\infty} E_i) = \sum_{k=1}^{\infty} P(E_i)$.

  (d) **Corollary 1: Complementation** $P(E) + P(E^C) = 1$

  (e) **Corollary 2: Monotonicity** If $E \subseteq F$, $P(E) \leq P(F)$

  (f) **Corollary 2: Inclusion-Exclusion** $P(E \cup F) = P(E) + P(F) - P(E \cap F)$

(k) **Equally Likely Outcomes:** If every outcome in a finite sample space S is equally likely, and E is an event, then $P(E) = \dfrac{|E|}{|S|}$.
  - Make sure to be consistent when counting $|E|$ and $|S|$. Either order matters in both, or order doesn't matter in both.

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

(m) **Independence**: Events E and F are independent iff $P(E \cap F) = P(E)P(F)$, or equivalently $P(F) = P(F|E)$, or equivalently $P(E) = P(E|F)$

(n) **Bayes Theorem**: $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$

(o) **Partition:** Nonempty events $E_1, \ldots, E_n$ partition the sample space S iff
  - $E_1, \ldots, E_n$ are exhaustive: $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^{n} E_i = S$, and
  - $E_1, \ldots, E_n$ are pairwise mutually exclusive: $\forall i \neq j, E_i \cap E_j = \emptyset$
    - Note that for any event A (with $A \neq \emptyset, A \neq S$): A and A^C partition S

(p) **Law of Total Probability (LTP)**: Suppose $A_1, \ldots, A_n$ partition S and let B be any event. Then $P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(\mathrm{B} \mid A_i)P(A_i)$

(q) **Bayes Theorem with LTP**: Suppose $A_1, \ldots, A_n$ partition S and let B be any event. Then $P(A_1|B) = \dfrac{P(B \mid A_1)P(A_1)}{\sum_{i=1}^{n} P(\mathrm{B} \mid A_i)P(A_i)}$. In particular, $P(A|B) = \dfrac{P(B \mid A)P(A)}{P(\mathrm{B} \mid A)P(A) + P(\mathrm{B} \mid A^C)P(A^C)}$

(r) **Chain Rule**: Suppose $A_1, ..., A_n$ are events. Then,

$$P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n|A_1 \cap ... \cap A_{n-1})$$

(s) **Conditional Independence**: Events E and F are conditionally independent given event G (with $P(G) > 0$) iff $P(E \cap F|G) = P(E|G)P(F|G)$, or equivalently $P(F|G) = P(F|E \cap G)$, or equivalently $P(E|G) = P(E|F \cap G)$

1. Discrete Random Variables

(a) **Random Variable (rv)**: A numeric function $X : \Omega \to \mathbb{R}$ of the outcome.

(b) **Range/Support:** The support/range of a random variable X, denoted Ω_X, is the set of all possible values that X can take on.

(c) **Discrete Random Variable (drv)**: A random variable taking on a countable (either finite or countably infinite) number of possible values.

(d) **Probability Mass Function (pmf) for a discrete random variable $\mathbf{X}$**: a function $p_X : \Omega_X \to [0, 1]$ with $p_X(x) = P(X = x)$ that maps possible values of a discrete random variable to the probability of that value happening, such that $\sum_x p_X(x) = 1$.

(e) **Expectation (expected value, mean, or average)**: The expectation of a discrete random variable is defined to be $\mathbb{E}[X] = \sum_x x p_X(x) = \sum_x x P(X = x)$. The expectation of a function of a discrete random variable $g(X)$ is $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$.

(f) **Linearity of Expectation**: Let X and Y be random variables, and $a, b, c \in \mathbb{R}$. Then, $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$.

(g) **Variance**: Let X be a random variable and $\mu = \mathbb{E}[X]$. The variance of X is defined to be $\text{Var}(X) = \mathbb{E}\left[(X - \mu)^2\right]$. Notice that since this is an expectation of a nonnegative random variable ($(X - \mu)^2$), variance is always nonnegative. With some algebra, we can simplify this to $\text{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

(h) **Standard Deviation**: Let X be a random variable. We define the standard deviation of X to be the square root of the variance, and denote it $\sigma = \sqrt{\text{Var}(X)}$.

(i) **Property of Variance**: Let $a, b \in \mathbb{R}$ and let X be a random variable. Then, $\text{Var}(aX + b) = a^2\text{Var}(X)$.

(j) **Independence**: Random variables X and Y are independent, written $X \perp Y$, iff

$$\forall x \forall y, \quad P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

In this case, we have $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ (the converse is not necessarily true).

(k) **i.i.d. (independent and identically distributed)**: Random variables $X_1, \ldots, X_n$ are i.i.d. (or iid) iff they are independent and have the same probability mass function.

(l) **Variance of Independent Variables**: If $X \perp Y$, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. This depends on independence, whereas linearity of expectation always holds. Note that this combined with the above shows that $\forall a, b, c \in \mathbb{R}$ and if $X \perp Y$, $\text{Var}(aX + bY + c) = a^2\text{Var}(X) + b^2\text{Var}(Y)$.

2. Zoo of Discrete Random Variables

(a) **Uniform**: $X \sim \text{Uniform}(a, b)$ ($\text{Unif}(a, b)$ for short), for integers $a \leq b$, iff X has the following probability mass function:

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

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

(b) **Bernoulli (or indicator)**: $X \sim \text{Bernoulli}(p)$ ($\text{Ber}(p)$ for short) iff X has the following probability mass function:

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

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

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

(c) **Binomial**: $X \sim \text{Binomial}(n, p)$ ($\text{Bin}(n, p)$ for short) iff X is the sum of n iid Bernoulli(p) random variables. X has probability mass function

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

$\mathbb{E}[X] = np$ and $\text{Var}(X) = np(1-p)$. An example of a Binomial r.v. is the number of heads in n independent flips of a coin with $P(\text{head}) = p$. Note that $\text{Bin}(1, p) \equiv \text{Ber}(p)$. As $n \to \infty$ and $p \to 0$, with $np = \lambda$, then $\text{Bin}(n, p) \to \text{Poi}(\lambda)$. If $X_1, \ldots, X_n$ are independent Binomial r.v.'s, where $X_i \sim \text{Bin}(N_i, p)$, then $X = X_1 + \ldots + X_n \sim \text{Bin}(N_1 + \ldots + N_n, p)$.

(d) **Geometric:** $X \sim \text{Geometric}(p)$ ($\text{Geo}(p)$ for short) iff X has the following probability mass function:

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

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

(e) **Negative Binomial**: $X \sim \text{NegativeBinomial}(r, p)$ ($\text{NegBin}(r, p)$ for short) iff X is the sum of r iid Geometric(p) random variables. X has probability mass function

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

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

(f) **Poisson**: $X \sim \text{Poisson}(\lambda)$ ($\text{Poi}(\lambda)$ for short) iff X has the following probability mass function:

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

$\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$. An example of a Poisson r.v. 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 r.v.'s, where $X_i \sim \text{Poi}(\lambda_i)$, then $X = X_1 + \ldots + X_n \sim \text{Poi}(\lambda_1 + \ldots + \lambda_n)$.

(g) **Hypergeometric**: $X \sim \text{HyperGeometric}(N, K, n)$ ($\text{HypGeo}(N, K, n)$ for short) iff X has the following probability mass function:

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

$\mathbb{E}[X] = n\frac{K}{N}$. 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 $\text{Bin}\left(n, \frac{K}{N}\right)$.

3. Continuous Random Variables

(a) **Cumulative Distribution Function (cdf)**: For any random variable (discrete or continuous) X, the cumulative distribution function is defined as $F_X(x) = P(X \leq x)$. Notice that this function must be monotonically nondecreasing: if $x < y$ then $F_X(x) \leq F_X(y)$, because $P(X \leq x) \leq P(X \leq y)$. Also notice that since probabilities are between 0 and 1, that $0 \leq F_X(x) \leq 1$ for all x, with $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.

(b) **Continuous Random Variable**: A continuous random variable X is one for which its cumulative distribution function $F_X(x) : \mathbb{R} \to \mathbb{R}$ is continuous everywhere. A continuous random variable has an uncountably infinite number of values.

(c) **Probability Density Function (pdf or density)**: Let X be a continuous random variable. Then the probability density function $f_X(x) : \mathbb{R} \to \mathbb{R}$ of X is defined as $f_X(x) = \frac{d}{dx} F_X(x)$. Turning this around, it means that $F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\, dt$. From this, it follows that $P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x) dx$ and that $\int_{-\infty}^{\infty} f_X(x) dx = 1$. From the fact that $F_X(x)$ is monotonically nondecreasing it follows that $f_X(x) \geq 0$ for every real number x.

If X is a continuous random variable, note that in general $f_X(a) \neq P(X = a)$, since $P(X = a) = F_X(a) - F_X(a) = 0$ for all a. However, the probability that X is close to a is proportional to $f_X(a)$: for small δ, $P\left(a - \frac{\delta}{2} < X < a + \frac{\delta}{2}\right) \approx \delta f_X(a)$.

(d) **i.i.d. (independent and identically distributed)**: Random variables $X_1, \ldots, X_n$ are i.i.d. (or iid) if they are independent and have the same probability mass function or probability density function.

(e) **Univariate: Discrete to Continuous:**

	Discrete	**Continuous**
PMF/PDF	$p_X(x) = P(X = x)$	$f_X(x) \neq P(X = x) = 0$
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	$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

(f) **Standardizing**: Let X be any random variable (discrete or continuous, not necessarily normal), with $\mathbb{E}[X] = \mu$ and Var$(X) = \sigma^2$. If we let $Y = \frac{X - \mu}{\sigma}$, then $\mathbb{E}[Y] = 0$ and Var$(Y) = 1$.

(g) **Closure of the Normal Distribution**: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. That is, linear transformations of normal random variables are still normal.

(h) **"Reproductive" Property of Normals**: Let $X_1, \ldots, X_n$ be independent normal random variables with $\mathbb{E}[X_i] = \mu_i$ and Var$(X_i) = \sigma_i^2$. Let $a_1, \ldots, a_n \in \mathbb{R}$ and $b \in \mathbb{R}$. Then,

$$X = \sum_{i=1}^{n} (a_i X_i + b) \sim \mathcal{N}\left(\sum_{i=1}^{n} (a_i \mu_i + b), \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right)$$

There's nothing special about the parameters – the important result here is that the resulting random variable is still normally distributed.

(i) **Law of Total Probability (Continuous)**: Let X, Y be continuous random variables. Then,

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

(j) **Convolution (Discrete)**: Let X, Y be independent discrete random variables, and $Z = X + Y$ be their convolution. Then,

$$p_Z(z) = P(X + Y = z) = \sum_x P(Y = z - X | X = x) P(X = x)$$

$$= \sum_x P(Y = z - x)P(X = x) = \sum_x p_X(x)p_Y(z - x)$$

(k) **Convolution (Continuous)**: Let X, Y be independent continuous random variables, and $Z = X + Y$ be their convolution. Then,

$$F_Z(z) = P(X + Y \leq z) = \int_{-\infty}^{\infty} P(Y \leq z - X | X = x) f_X(x)dx = \int_{-\infty}^{\infty} F_Y(z - x) f_X(x)dx$$

Hence, $f_Z(z) = F_Z'(z) = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x)dx$.

(l) **Multivariate: Discrete to Continuous:**

	Discrete	**Continuous**
Joint PMF/PDF	$p_{X,Y}(x, y) = P(X = x, Y = y)$	$f_{X,Y}(x, y) \neq P(X = x, Y = y)$
Joint CDF	$F_{X,Y}(x, y) = \sum_{t \leq x, s \leq y} p_{X,Y}(t, s)$	$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t, s)\, dsdt$
Normalization	$\sum_{x,y} p_{X,Y}(x, y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dxdy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_y p_{X,Y}(x, y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$
Expectation	$\mathbb{E}[g(X, Y)] = \sum_{x,y} g(x, y)p_{X,Y}(x, y)$	$\mathbb{E}[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)f_{X,Y}(x, y)dxdy$
Conditional PMF/PDF	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$\mathbb{E}[X \mid Y = y] = \sum_x x p_{X\|Y}(x\|y)$	$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X\|Y}(x\|y)dx$
Independence	$\forall x, y, p_{X,Y}(x, y) = p_X(x)p_Y(y)$	$\forall x, y, f_{X,Y}(x, y) = f_X(x)f_Y(y)$

(m) **Law of Total Expectation**: Let X be a random variable (discrete or continuous). If Y is a discrete random variable, then

$$\mathbb{E}[X] = \sum_y \mathbb{E}[X \mid Y = y]\, p_Y(y)$$

If Y is a continuous random variable, then

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

4. Zoo of Continuous Random Variables

(a) **Uniform**: $X \sim \text{Uniform}(a, b)$ iff X has the following probability density function:

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

$\mathbb{E}[X] = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$. This represents each real number from $[a, b]$ to be equally likely.

(b) **Exponential**: $X \sim \text{Exponential}(\lambda)$ iff X has the following probability density function:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$. $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The exponential random variable is the continuous analog of the geometric random variable: 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 random variable X is memoryless:

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

The geometric random variable also has this property.

(c) **Normal (Gaussian, "bell curve")**: $X \sim \mathcal{N}(\mu,\ \sigma^2)$ iff X has the following probability density function:

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

$\mathbb{E}[X] = \mu$ and $\mathsf{Var}(X) = \sigma^2$. 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 $\Phi\left(z\right) = F_Z\left(z\right) = P(Z \leq z)$. Note from symmetry of the probability density function about $z = 0$ that: $\Phi\left(-z\right) = 1 - \Phi(z)$.