CSE 312: Foundations of Computing II **Summer 2020**

Cheat Sheet

Lecturer: Alex Tsun *Scribe: Mitchell Estberg*

0 Prerequisites

Sets and Cardinality

Definition 0.1.1: Set

A **set**, S for example, is an unordered collection of objects (with no duplicates), finite or infinite.

Definition 0.1.2: Cardinality

The **cardinality** of S is denoted $|S|$, which is the number of elements in the set.

Definition 0.1.3: Empty Set

There is only onset set of cardinality 0, the **empty set**, denoted by $\emptyset = \{\}$

Subsets and Equality

Definition 0.1.4: In and Not In

If x is in a set S, we write $x \in S$, If x is not in set S, we write $x \notin S$.

Definition 0.1.5: Subset

We write $A \subset B$ to mean A is a **subset** of B, that is for any $x \in A$, it must be the case that $x \in B$.

Definition 0.1.6: Superset

We write $A \supset B$ to mean that A is a **superset** of B (equiavlent to $B \subset A$).

Definition 0.1.7: Set Equality

We say two sets A, B ae equal ($A = B$) if and only if $A \subset B$ and $B \subset A$.

Set Operations

Definition 0.2.8: Universal Set

Let A, B be sets and U be a **unversal set**, so that $A \subset U$ and $B \subset U$.

Definition 0.2.9: Set Operation: Union

The **union** of A and B is denoted $A \cup B$. It contains elements in A or B, or both (without duplicates). So $x \in A \cup B$ if and only if $x \in A$ or $x \in B$.

Definition 0.2.10: Set Operation: Intersection

The **intersection** of A and B is denoted $A \cap B$. It contains elements in A and B. So $x \in A \cap B$ if and only if $x \in A$ and $x \in B$.

Definition 0.2.11: Set Operation: Set Difference

The **set difference** of A with B is denoted, $A \setminus B$. It contains elements of A which are not in B. So $x \in A \setminus B$ if and only if $x \in A$ and $x \notin B$.

Definition 0.2.12: Set Operation: Complement

The **complement** with respect to U of A is denoted $A^C = U \setminus A$. It contains elements of U, the universal set, which are not in A

Summation Notation

Definition 0.3.13: Summation Notation

Let $x_1, x_2, x_3, \ldots$ be a sequence of numbers. Then, the following notation represents the sum $x_a + x_{a+1} + \cdots + x_{b-1} + x_b$: $\sum_{i=a}^{b} x_i$. Further, if S is a set, and $f : S \to \mathbb{R}$ is a function defined on S, then the following notation sums over all elements $x \in S$ of $f(x)$: $\sum_{x \in S} f(x)$. Note that the sum over no terms is defined as 0.

Fact 0.3.1: The Associative and Distributive Properties of Sums

1. $\sum_{x \in A} f(x) + \sum_{x \in A} g(x) = \sum_{x \in A}(f(x) + g(x))$
2. $\sum_{x \in A} \alpha \cdot f(x) = \alpha \sum_{x \in A}(f(x))$
3. $(\sum_{x \in A} f(x))(\sum_{y \in B} g(x)) = \sum_{x \in A} \sum_{y \in b} f(x)g(x)$

Product Notation

Definition 0.3.14: Product Notation

Let $x_1, x_2, x_3, \ldots$ be a sequence of numbers. Then, the following notation represents the sum $x_a \cdot x_{a+1} \cdot \cdots \cdot x_{b-1} \cdot x_b$: $\prod_{i=a}^{b} x_i$. Further, if S is a set, and $f : S \to \mathbb{R}$ is a function defined on S, then the following notation multiplies over all elements $x \in S$ of $f(x)$: $\prod_{x \in S} f(x)$. Note that the product over no terms is defined as 1.

1 Counting

Sum Rule

Definition 1.1.15: 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 number of possible outcomes of the experiment is $N + M$.

Product Rule

Definition 1.1.16: 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 $N_1 \times N_2 \times \cdots \times N_m$.

Permutations

Definition 1.1.17: Permutation

The number of orderings of N **distinct** objects, is called a permutation, and mathematically defined as: $N! = N \times (N-1) \times (N-2) \times \ldots 3 \times 2 \times 1$, with $N!$ pronounced "N factorial".

Complementary Counting

Definition 1.1.18: Complementary Counting

Let U be a (finite) universal set, and S a subset of interest. Let $U \setminus S$ denote the set difference. Then, $\mid S \mid = \mid U \mid - \mid U \setminus S \mid$ That is, the complement of the subset of interest is also of interest!

k-Permutations

Definition 1.2.19: k-Permutations

If we want to arrange **only** k out of n distinct objects, the number of ways to do so is: $P(n, k) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$ (Read as "n pick k"). A **permutation** of a set is an arrangement of its members where order matters, so a **k-permutation** is the arrangement of k members of a set of n members where order matters.

Combinations/Binomial Coefficients

Definition 1.2.20: Combinations/Binomial Coefficients

If we want to select (order doesn't matter) **only** k out of n distinct objects, the number of ways to do so is: $C(n, k) = \binom{n}{k} = \frac{P(n,k)}{k!} = \frac{n!}{k!(n-k)!}$. A **combination** is a selection of items from a set in which the order does not matter.

Multinomial Coefficients

Definition 1.2.21: Multinomial Coefficients

If we have k types of objects (n total), with n_1 of the first type, n_2 of the second, ..., and n_k of the kth, 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

Definition 1.2.22: Stars and Bars/Divider Method

The number of ways to distribute n indistinguishable balls into k distinguishable bins is: $\binom{n+(k-1)}{k-1} = \binom{n+(k-1)}{n}$.

Binomial Theorem

Theorem 1.3.1: Binomial Theorem

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

Inclusion-Exclusion

Theorem 1.3.2: Inclusion-Exclusion

Let A, B be sets, then: $|A \cup B| = |A| + |B| - |A \cap B|$ Further, in general, if $A_1, A_2, \ldots, A_n$ are sets, then: $|A_1 \cup \cdots \cup A_n| = \text{singles} - \text{doubles} + \text{triples} - \text{quads} + \cdots = (|A_1| + \cdots + |A_n|) - (|A_1 \cap A_2| + \cdots + |A_{n-1} \cap A_n|) + (A_1 \cap A_2 \cap A_3| + \cdots + |A_{n-2} \cap A_{n-1} \cap A_n|) + \cdots$. Where singles are the size of all the single sets, doubles are the size of all the intersections of two sets, triples are the size of all the intersections of three sets, quads are all the intersection of four sets, and so forth.

Pigeonhole Principle

Definition 1.3.23: Floor and Ceiling Functions

The **floor** function $\lfloor x\ rfloor$ returns the largest integer $\leq x$ (i.e. rounds down).
The **ceiling** function $\lceil x \rceil$ returns the smallest integer $\geq x$ (i.e. rounds up).

Theorem 1.3.3: 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 pigeons. More generally, if there are n pigeons we want to put into k pigeonholds, then at least one pigeonhold must contain at least $\lceil n/k \rceil$ pigeons.

Combinatorial Proofs

Definition 1.3.24: 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

Definitions for Probability

Definition 2.1.25: Sample Space

The **sample space** is the set Ω of all possible outcomes of an experiment.

Definition 2.1.26: Event

An **event** is any subset $E \subseteq \Omega$

Definition 2.1.27: Mutually Exclusision

Events E and F are considered mutually exclusive if $E \cap F = \emptyset$. (they can't simultaneously happen).

Axioms of Probability and Consequences

Definition 2.1.28: Axioms of Probability

Let Ω denote the sample space and $E, F \subseteq \Omega$ be events.

1. (Nonnegativity) $\mathbb{P}(E) \geq 0$, that is no event has a negative probability.
2. (Normalization) $\mathbb{P}(\Omega) = 1$, that is the probability of the entire sample space is always 1. (Something is guaranteed to happen)
3. (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

Theorem 2.1.4: Probability in Sample Space with 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 $E \subseteq \Omega$: $\mathbb{P}(E) = \dfrac{|E|}{|\Omega|}$.

Conditional Probability

Definition 2.2.29: Conditional Probability

$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, or in other words: $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$

Bayes Theorem

Theorem 2.2.5: Bayes Theorem

Let A, B be events with nonzero probability. Then, $\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$

Law of Total Probability

Definition 2.2.30: Partitions

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

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

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

Theorem 2.2.6: Law of Total Probability (LTP)

If events $E_1, \ldots, E_n$ partition Ω, then for any event F, then $\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \cdots + \mathbb{P}(F \cap E_n) = \sum_{i=1}^n \mathbb{P}(F \cap E_i)$, or equivalently (by definition of conditional probability),
$\mathbb{P}(F) = \mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1) + \cdots + \mathbb{P}(F \mid E_n)\,\mathbb{P}(E_n) = \sum_{i=1}^n \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)$.

Bayes Theorem with the Law of Total Probability

Theorem 2.2.7: Bayes Theorem with the Law of Total Probability

Let events $E_1, \ldots, E_n$ partition the sample space Ω, and let F be another event. Then,
$\mathbb{P}(E_1 \mid F) = \frac{\mathbb{P}(F|E_1)\mathbb{P}(E_1)}{\sum_{i=1}^n \mathbb{P}(F|E_i)\mathbb{P}(E_i)}$. In particular, in the case of a simple partition of Ω into E and E^C, if E is an event with nonzero probability, then, $\mathbb{P}(E \mid F) = \frac{\mathbb{P}(F|E)\mathbb{P}(E)}{\mathbb{P}(F|E)\mathbb{P}(E) + \mathbb{P}(F|E^C)\mathbb{P}(E^C)}$.

Chain Rule

Theorem 2.3.8: Chain Rule

For $A_1, \ldots, A_n$ be events with nonzero probabilities. Then:
$\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 A_2) \cdots \mathbb{P}(A_n \mid A_1, \ldots, A_{n-1})$.
In the case of two events, A, B: $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)$.

Independence

Definition 2.3.31: Independence

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

Conditional Independence

Definition 2.3.32: Conditional Independence

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

Notice that this is very similar to the definition of independence. There is no difference, except we have just added in conditioning on C to every probability.

3 Discrete RVs

Introduction to Discrete Random Variables

Definition 3.1.33: Random Variable

Suppose we conduct an experiment with sample space Ω. A **random variable (rv)** is a numeric function of the outcome, $X : \Omega \to \mathbb{R}$. That is it maps outcomes $\omega \in \Omega$ to numbers, $\omega \to X(\omega)$.
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 random variable (drv)**. Else if Ω_X is uncountably large (the size of real numbers), X is a **continuous random variable**.

Probability Mass Functions (PMFs)

Definition 3.1.34: Probability Mass Function (pmf)

The **probability mass function (pmf)** of a discrete random variable X assigns probabilities to the possible values of the random variable. That is $p_X : \Omega_X \to [0, 1]$ where:

$$p_X(k) = \mathbb{P}(X = k)$$

Note that $\{X = a\}$ for $a \in \Omega$ form a partition of Ω, since each outcome $\omega \in \Omega$ is mapped to exactly one number. Hence,

$$\sum_{z \in \Omega_X} p_X(z) = 1$$

Notice here the only thing consistent is p_X, as it's the PMF of X. The value inside is a dummy variable - just like we can write $f(x) = x^2$ or $f(t) = t^2$. To reinforce this, I will constantly use different letters for dummy variables.

Expectation

Definition 3.1.35: Expectation

The **expectation/expected value/average** of a discrete random variable X is:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$

or equivalently,

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k)$$

The interpretation is that we take an average of the values in Ω_X, but weighted by their probabilities.

Linearity of Expectation

Theorem 3.2.9: Linearity of Expectation (LoE)

Let Ω be the sample space of an experiment, $X, Y : \Omega \to \mathbb{R}$ be (possibly "dependent") random variables both defined on Ω, and $a, b, c \in \mathbb{R}$ be scalars. Then,

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

and

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

Combining them gives,

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

Law of the Unconcious Statistician

Theorem 3.2.10: Law of the Unconscious Statistician (LOTUS)

Let X be a discrete random variable with range Ω_X and $g : D \to \mathbb{R}$ be a function defined at least over $\Omega_X, (\Omega_X \subseteq D)$. Then

$$E[g(X)] = \sum_{b \in \Omega_X} g(b) p_X(b)$$

Note that in general, $\mathbb{E}\left[g(X)\right] \neq g(\mathbb{E}\left[X\right])$. For example, $\mathbb{E}\left[X^2\right] \neq (\mathbb{E}\left[X\right])^2$, or $\mathbb{E}\left[\log(X)\right] \neq \log(\mathbb{E}\left[X\right])$

Linearity of Expectation with RVs

Claim 3.3.1: Linearity of Expectation with Indicators

If asked only about the expectation of a random variable X (and not its PMF), then you may be able to write X as the sum of possibly dependent indicator random variables, and apply linearity of expectation.

For an indicator random variable X_i,

$$\mathbb{E}\left[X_i\right] = 1 \cdot \mathbb{P}\left(X_i = 1\right) + 0 \cdot \mathbb{P}\left(X_i = 0\right) = \mathbb{P}\left(X_i = 1\right)$$

Variance

Definition 3.3.36: Variance

The variance of a random variable X is defined to be

$$Var(X) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$$

The variance is always nonnegative since we take the expectation of a nonnegative random variable $(X - \mathbb{E}\left[X\right])^2$. The first equality is the *definition* of variance, and the second equality is a more useful identity which we will need to prove.

Definition 3.3.37: Standard Deviation

Another measure of a random variable X's spread is the standard deviation, which is

$$\sigma_X = \sqrt{Var(X)}$$

This measure is also useful, because the units of variance are squared in terms of the original variable X, and this essentially "undos" our squaring, returning our units to the same as X.

Corollary 3.3.1: Property of Variance

We can also show that for any scalar $a, b \in \mathbb{R}$,

$$Var(aX + b) = a^2 Var(X)$$

Independence of Random Variables

Definition 3.4.38: Independence

Random variables X and Y are independent, denoted $X \perp Y$, if for all $x \in \Omega_X$ and all $y \in \Omega_Y$, any of the following three equivalent properties holds:

1. $P(X = x | Y = y) = P(X = x)$
2. $P(Y = y | X = x) = P(Y = y)$
3. $P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$

Note, that this is the same as the event definition of independence, but it must hold for all events $\{X = x\}$ and $\{Y = y\}$.

Fact 3.4.2: Variance Adds for Independent RVs

If $X \perp Y$, then

$$Var(X + Y) = Var(X) + Var(Y)$$

This will be proved a bit later, but we can start using this fact now!

A common misconception is that $Var(X - Y) = Var(X) - Var(Y)$, but this actually isn't true, otherwise we could get a negative number. In fact, if $X \perp Y$, then

$$Var(X-Y) = Var(X+(-Y)) = Var(X)+Var(-Y) = Var(X)+(-1)^2 Var(Y) = Var(X)+Var(Y)$$

The Bernoulli Process and Bernoulli Random Variable

Definition 3.4.39: Bernoulli Process

A Bernoulli process with parameter p is a sequence of independent coin flips $X_1, X_2, X_3, ...$ where $P(head) = p$. If flip i is heads, then we encode $X_i = 1$; otherwise, $X_i = 0$. From this process we can measure many interesting things.

Definition 3.4.40: Bernoulli/Indicator Random Variable

A random variable X is Bernoulli (or indicator), denoted $X \sim Ber(p)$, if and only if X has the following PMF:

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

Each X_i in the Bernoulli process with parameter p is Bernoulli/indicator random variable with parameter p. It simply represents a binary outcome, like a coin flip.

Additionally,

$$\mathbb{E}[X] = p \text{ and } Var(X) = p(1-p)$$

The Binomial Random Variable

Definition 3.4.41: Binomial Random Variable

A random variable X has a Binomial distribution, denoted $X \sim Bin(n, p)$, if and only if X has the following PMF:

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

X is the sum of n independent $Ber(p)$ random variables, and represents the number of heads in n independent coin flips where $P(head) = p$.

Additionally,

$$E[X] = np \text{ and } Var(X) = np(1-p)$$

The Uniform (Discrete) Random Variable

Definition 3.5.42: Uniform Random Variable

X is a uniform random variable, denoted $X \sim Unif(a, b)$, where $a < b$ are integers, if and only if X has the following probability mass function

$$p_X(k) \begin{cases} \frac{1}{b-a+1}, & k \in \{a, a+1, ..., b\} \\ 0, & \text{otherwise} \end{cases}$$

X is equally likely to take on any value in $\Omega_X = \{a, a+1, ..., b\}$. This set contains $b - a + 1$ integers, which is why $\mathbb{P}(X = k)$ is always $\frac{1}{b-a+1}$.

Additionally,

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

As you might expect, the expected value is just the average of the endpoints that the uniform random variable is defined over.

The Geometric Random Variable

Definition 3.5.43: Geometric Random Variable

X is a Geometric random variable, denoted $X \sim Geo(p)$, if and only if X has the following probability mass function (and range $\Omega_X = \{1, 2, \dots\}$):

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

Additionally,

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

The Negative Binomial Random Variable

Definition 3.5.44: Negative Binomial Random Variable

X is a negative binomial random variable, denoted $X \sim NegBin(r, p)$, if and only if X has the following probability mass function (and range $\Omega_X = \{r, r+1, \dots,\}$):

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

X is the sum of r independent $Geo(p)$ random variables.

Additionally,

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

Also, note that $Geo(p) \equiv NegBin(1, p)$, and that if X, Y are independent such that $X \sim NegBin(r, p)$ and $Y \sim NegBin(s, p)$, then $X + Y \sim NegBin(r + s, p)$ (waiting for $r + s$ heads).

The Poisson Random Variable

Definition 3.6.45: The Poisson Variable PMF

Let λ be the historical average number of events per unit of time. Send $n \to \infty$ in such a way that $np = \lambda$ is fixed (i.e., $p = \frac{\lambda}{n}$).

Let $X_n \sim Bin(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n\to\infty} X_n$ by the limit of this sequence of Binomial rvs. Then, we say $Y \sim Poi(\lambda)$ and measures the number of events in a unit time, where the historical average is λ, and has PMF

$$p_Y(k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Fact 3.6.3

The Poisson RV PMF sums to 1.

Lemma 3.6.1: Poisson RV properties

Let $X_n \sim Bin(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n\to\infty} X_n = Poi(\lambda)$.
By the properties of the binomial random variable:

$$\mathbb{E}[X_n] = np = \lambda$$

$$Var(X_n) = np(1-p) = \lambda \left(1 - \frac{\lambda}{n}\right)$$

Therefore:

$$\mathbb{E}[Y] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}[X_n] = \lim_{n\to\infty} \lambda = \lambda$$

$$Var(Y) = Var\left(\lim_{n\to\infty} X_n\right) = \lim_{n\to\infty} Var(X_n) = \lim_{n\to\infty} \lambda \left(1 - \frac{\lambda}{n}\right) = \lambda$$

Definition 3.6.46: The Poisson RV

$X \sim Poi(\lambda)$ if and only if X has the following pmf (and range $\Omega_X = \{0, 1, 2, \dots\}$):

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, k = 0, 1, 2, ...$$

If λ is the historical average of events per unit of time, then X is the number of events that occur in a unit of time.
We also computed earlier that

$$\mathbb{E}[X] = \lambda, \qquad Var(X) = \lambda$$

The Poisson Process

Definition 3.6.47: The Poisson Process

A Poisson process with rate $\lambda > 0$ per unit of time, is a continuous-time stochastic process indexed by $t \in [0, \infty)$, so that $X(t)$ is the number of events that happens in the interval $[0, t]$. Notice that if $t_1 < t_2$, then $X(t_2) - X(t_1)$ is the number of events in $(t_1, t_2]$. The process has three properties:

- $X(0) = 0$. That is, we initially start with an empty counter at time 0.
- The number of events happening in any two disjoint intervals $[a, b]$ and $[c, d]$ are independent.
- The number of events in any time interval $[t_1, t_2]$ is $Poi(\lambda(t_2 - t_1))$. This is because on average λ events happen per unit time, so in $t_2 - t_1$ units of time, the average rate is $\lambda(t_2 - t1)$. Again, we can scale our rate but not our period of interest.

The Hypergeometric Random Variable

Definition 3.6.48: The Hypergeometric RV

$X \sim HypGeo(N, K, n)$ if and only if X has the following pmf:

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

Lemma 3.6.2: Hypergeometric RV properties

Suppose $X \sim HypGeo(N, K, n)$, let $X_1, ..., X_n$ be indicator RV's (not independent) so that $X_i = 1$ if we got a lollipop on the i^{th} draw, and 0 otherwise. So $X = \sum_{i=1}^{n} X_i$.

Then, each X_i is Bernoulli, but with what parameter?

$$\mathbb{P}(X_1 = 1) = \frac{K}{N}$$

$$\mathbb{P}(X_2 = 1) = \mathbb{P}(X_2 = 1|X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{P}(X_2 = 1|X_1 = 0)\mathbb{P}(X_1 = 0) \qquad \text{[LTP]}$$

$$= \frac{K-1}{N-1}\cdot\frac{K}{N} + \frac{K}{N-1}\cdot\frac{N-K}{N} = \frac{K(N-1)}{N(N-1)} = \frac{K}{N}$$

Actually, each $X_i \sim Ber(K/N)$ independent of i! You could continue the above logic for X_3 and so on.

$$\mathbb{E}[X_i] = p = \frac{K}{N}$$

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{K}{N} = n\frac{K}{N}$$

Note again it would be wrong to say $X \sim Bin(n, K/N)$ because the trials are NOT independent, but we are still able to use linearity of expectation.

4 Continuous RVs

Definition 4.1.49: Continuous Random Variables

A continuous random variable is a random variable that takes values from an uncountable, infinite set, such as the set of real numbers. For e.g., height (5.6312435 feet, 6.1123 feet, etc.), weight (121.33567 lbs, 153.4642 lbs, etc.) and time (2.5644 seconds, 9321.23403 seconds, etc.) are continuous random variables that take on values in a continuum.

Probability Density Functions (PDFs)

Definition 4.1.50: Probability Density Function (PDF)

Let X be a continuous random variable (one whose range is typically an interval or union of intervals). The probability density function (PDF) of X is the function $f_X : \mathbb{R} \to \mathbb{R}$, such that the following properties hold:

- $f_X(z) \geq 0$ for all $z \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(w)\, dw$
- $\mathbb{P}(X = y) = 0$ for any $y \in \mathbb{R}$
- The probability that X is close to q is proportional to its density $f_X(q)$;

$$\mathbb{P}(X \approx q) = \mathbb{P}\left(q - \frac{\varepsilon}{2} \leq X \leq q + \frac{\varepsilon}{2}\right) \approx \varepsilon f_X(q)$$

- Ratios of probabilities of being “near points” are maintained;

$$\frac{\mathbb{P}(X \approx u)}{\mathbb{P}(X \approx v)} \approx \frac{\varepsilon f_X(u)}{\varepsilon f_X(v)} = \frac{f_X(u)}{f_X(v)}$$

Cumulative Distribution Functions (CDFs)

Definition 4.1.51: Cumulative Distribution Function (CDF)

Let X be a continuous random variable (one whose range is typically an interval or union of intervals). The cumulative distribution function (CDF) of X is the function $f_X : \mathbb{R} \to \mathbb{R}$ such that:

- $F_X(t) = P(X \leq t) = \int_{-\infty}^{t} f_X(w)\, dw$ for all $t \in \mathbb{R}$
- $\frac{d}{du} F_X(u) = f_X(u)$
- $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$
- F_X is monotone increasing, since $f_X \geq 0$. That is, $F_X(c) \leq F_X(d)$ for $c \leq d$.
- $\lim_{v \to -\infty} F_X(v) = \mathbb{P}(X \leq -\infty) = 0$
- $\lim_{v \to +\infty} F_X(v) = \mathbb{P}(X \leq +\infty) = 1$

From Discrete to Continous

	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/LOTUS	$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$

The (Continuous) Uniform RV

Definition 4.2.52: Uniform (Continuous) RV

$X \sim Unif(a, b)$ where $a < b$ are real numbers, if and only if X has the following pdf:

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

X is equally likely to be take on any value in $[a, b]$. Note the similarities and differences it has with the discrete uniform!

$$\mathbb{E}[X] = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12}$$

The cdf is

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \leq x \leq b \\ 1, & x > b \end{cases}$$

The Exponential RV

Claim 4.2.2: The Exponential RV properties

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

$$Var(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Definition 4.2.53: The Exponential RV

$X \sim Exp(\lambda)$, if and only if X has the following pdf (and range $\Omega_X = [0, \infty)$):

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

X is the waiting time until the first occurrence of an event in a Poisson Process with parameter λ.

$$\mathbb{E}[X] = \frac{1}{\lambda}, Var(X) = \frac{1}{\lambda^2}$$

The cdf is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Memorylessness

Definition 4.2.54: Memorylessness

A random variable X is **memoryless** is for all $s, t \geq 0$,

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$$

Theorem 4.2.11: Memorylessness of Exponential

If $X \sim Exp(\lambda)$, then X has the memoryless property.

The Gamma RV

Definition 4.2.55: Gamma RV

$X \sim Gamma(r, \lambda)$ if and only if X has the following pdf:

$$f_X(x) = \begin{cases} \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

X is the sum of r independent $Exp(\lambda)$ random variables.
Gamma is to exponential as negative binomial to geometric. It is the waiting time until the r^{th} event,

rather than just the first event. So you can write it as a sum of independent exponential random variables.

$$\mathbb{E}[X] = \frac{r}{\lambda}, Var(X) = \frac{r}{\lambda^2}$$

X is the waiting time until the r^{th} occurrence of an event in a Poisson Process with parameter λ. Notice that $Gamma(1, \lambda) \equiv Exp(\lambda)$. By definition, if X, Y are independent with $X \sim Gamma(r, \lambda)$ and $Y \sim Gamma(s, \lambda)$, then $X + Y \sim Gamma(r + s, \lambda)$.

The Normal/Gaussian Random Variable

Definition 4.3.56: Normal (Gaussian, "bell curve") distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if X has the following PDF (and range $\Omega_X = (-\infty, +\infty)$):

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where $\exp(y) = e^y$. This Normal random variable actually has as parameters its mean and variance, and hence:

$$\mathbb{E}[X] = \mu$$

$$Var(X) = \sigma^2$$

Closure Properties of the Normal Random Variable

Fact 4.3.4: 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,

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Fact 4.3.5: Closure of the Normal Under Addition

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ (both independent normal random variables), then

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

The Standard Normal CDF

Definition 4.3.57: Standard Normal Random Variable

The "standard normal" random variable is typically denoted Z and has mean 0 and variance 1. By the closure property of normals, 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 by $\Phi(a) = F_Z(a) = P(Z \leq a)$. Note that by symmetry of the density about 0, $\Phi(-a) = 1 - \Phi(a)$.

Transforming 1-D RVs via CDF

Definition 4.4.58: Steps to get PDF of $Y = g(X)$ from X (via CDF)

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, $F_Y(y) = P(g(X) \leq y)$, in terms of F_X.
4. Differentiate the CDF $F_Y(y)$ to get the PDF $f_Y(y)$ on Ω_Y. f_Y is 0 outside Ω_Y.

Transforming 1-D RVs via Explicit Formula

Theorem 4.4.12: Formula to get PDF of $Y = g(X)$ from X

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

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

Transforming Multidimensional RVs via Formula

Definition 4.4.59: Formula to get PDF of $Y = g(X)$ from X (Multidimensional Case)

Let $X = (X_1, ..., X_n)$, $Y = (Y_1, ..., Y_n)$ be continuous random vectors (each component is a continuous rv) with the same dimension n (so $\Omega_X, \Omega_Y \subseteq \mathbb{R}^n$), and $Y = g(X)$ where $g : \Omega_X \to \Omega_Y$ is invertible and differentiable, with differentiable inverse $X = g^{-1}(y) = h(y)$. Then,

$$f_Y(y) = f_X(h(y)) \left| \det\left(\frac{\partial h(y)}{\partial y}\right) \right|$$

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

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

5 Multiple RVs

Cartesian Product of Sets

Definition 5.1.60: Cartesian Product of Sets

Let A, B be sets. The Cartesian product of A and B is denoted:

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

Further if A, B are finite sets, then $|A \times B| = |A| \cdot |B|$ by the product rule of counting.

Joint PMFs and Expectation

Definition 5.1.61: 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

Definition 5.1.62: Marginal PMFs

Let X, Y be discrete random variables. The marginal PMF of X is:

$$p_X(a) = \sum_{b\in\Omega_Y} p_{X,Y}(a,b)$$

Similarly, the marginal PMF of Y is:

$$p_Y(d) = \sum_{c\in\Omega_X} p_{X,Y}(c,d)$$

(Extension) If Z is also a discrete random variable, then the marginal PMF of z is:

$$p_Z(z) = \sum_{x\in\Omega_X} \sum_{y\in\Omega_Y} p_{X,Y,Z}(x,y,z)$$

Independence

Definition 5.1.63: Independence (DRVs)

Discrete random variables X, Y are independent, written $X \perp Y$, if for all $x \in \Omega_X$ and $y \in \Omega_Y$:

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

Fact 5.1.6: Check for Independence (DRVs)

Recall $\Omega_{X,Y} = \{(x,y) : p_{X,Y}(x,y) > 0\} \subseteq \Omega_X \times \Omega_Y$. A necessary but not sufficient condition for independence is that $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. That is, if $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$, then X and Y cannot be independent, but if $\Omega_{X,Y} = \Omega_X \times \Omega_Y$, then we have to check the condition above.
This is because if there is some $(a,b) \in \Omega_X \times \Omega_Y$ but not in $\Omega_{X,Y}$, then $p_{X,Y}(a,b) = 0$ but $p_X(a) > 0$ and $p_Y(b) > 0$, violating independence. For example, suppose the joint PMF looks like:

$X \setminus Y$	8	9	Row Total
3	1/3	1/2	5/6
7	1/6	0	1/6
Col Total	1/2	1/2	1

Also side note that the marginal distributions are named what they are, since we often write the row and column totals in the margins. The joint range $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$ since one of the entries is 0, and so $(7,9) \notin \Omega_{X,Y}$ but $(7,9) \in \Omega_X \times \Omega_Y$. This immediately tells us they cannot be independent - $p_X(7) > 0$ and $p_Y(9) > 0$, yet $p_{X,Y}(7,9) = 0$.

Variance Adds for Independent Random Variables

Lemma 5.1.3: Variance Adds for Independent RVs

If X, Y are independent random variables, denoted $X \perp Y$, then:

$$Var(X+Y) = Var(X) + Var(Y)$$

If $a, b, c \in R$ are scalars, then:

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y)$$

Note this property relies on the fact that they are independent, whereas linearity of expectation always holds, regardless.

Joint Continous Distrbutions

Definition 5.2.64: Joint Continuous Distributions of two RVs

If two random variables X and Y are jointly continuous, then there exist a joint PDF $f_{X,Y}$ defined over $-\infty < x, y < \infty$ such that:

$$P(a_1 \leq X < a_2, b_1 \leq Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x,y)dydx$$

Two requirements must be satisfied for all continuous distributions:

1. $f_{X,Y}(x,y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx = 1$

Definition 5.2.65: Joint Range of two continuous RVs

The joint range of two continuous random variables X and Y is:

$$\Omega_{X,Y} = \{(x, y) : f_{X,Y}(x, y) \geq 0\} \subseteq \Omega_X \times \Omega_Y$$

Expectation of Jointly Distributed Random Variables

Definition 5.2.66: Expectation of Functions of Jointly Distributed Continuous Random Variables

Suppose that X and Y are jointly distributed continuous random variables with joint PDF $f(x, y)$. If $g(x, y) : \mathbb{R}^{\nvDash} \to \mathbb{R}$ is a function of these two random variables, then its expected value is given by the following:

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

Marginal PDFs

Definition 5.2.67: Marginal PDFs

Suppose that X and Y are jointly distributed continuous random variables with joint PDF $f(x, y)$. The marginal PDFs of X and Y are respectively given by the following:

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

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

(Extension): If Z is also a continuous random variable, then the marginal PDF of Z is:

$$f_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x, y, z) dx dy$$

Definition 5.2.68: Expected value (for jointly continuous random variables)

If we write the marginal $f_X(x)$ and $f_Y(y)$ in terms of joint density, then the expected values of X and Y are:

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

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

Independence of Continuous Random Variables

Definition 5.2.69: Independence of Continuous Random Variables

Continuous random variables X, Y are independent, written $X \perp Y$, if for all $x \in \Omega_X$ and $y \in \Omega_Y$,

$$f_{X,Y}(x,y) = f_X(x) f_Y(y)$$

Recall $\Omega_{X,Y} = \{(x,y) : f_{X,Y}(x,y) > 0\} \subseteq \Omega_Y \times \Omega_Y$. A necessary but not sufficient condition for independence is that $\Omega_{X,Y} = \Omega_X \times \Omega Y$. That is, if $\Omega_{X,Y} = \Omega_X \times \Omega_Y$, then we have to check the condition.
This is because if there is some $(a,b) \in \Omega_X \Omega_Y$ but not in $\Omega_{X,Y}$, then $f_{X,Y}(a,b) = 0$ but $f_X(a) > 0$ and $f_Y(b) > 0$, which violates independence.

Multivariate: From 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	$F_{X,Y}(x,y) = \sum_{t \le x} \sum_{s \le y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s) ds dt$
Normalization	$\sum_x \sum_y p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$
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	$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(x,y)$	$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy$
Conditional PMF/PDF	$p_{X \mid Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$E[X \mid Y = y] = \sum_x x p_{X \mid Y}(x \mid y)$	$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X \mid Y}(x \mid y) dx$
Independence	$\forall x, y, p_{X,Y}(x,y) = p_X(x) p_Y(y)$	$\forall x, y, f_{X,Y}(x,y) = f_X(x) f_Y(y)$

Conditional PMFs and PDFs

Definition 5.3.70: Conditional PMFs and PDFs

If X, Y are discrete random variables, then the conditional PMF of X given Y is:

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

Note that the final step is by Bayes Theorem.
If X, Y are continuous random variables, then the conditional PDF of X given Y is:

$$f_{X|Y}(u \mid v) = \frac{f_{X,Y}(u,v)}{f_Y(v)} = \frac{f_{Y|X}(v \mid u) f_X(u)}{f_Y(v)}$$

If X and Y are mixed (one discrete, one continuous), then a similar extension can be made where any discrete random variable has a p (a probability mass function) any continuous random variable has an f (a probability density function).

Conditional Expectation

Definition 5.3.71: Conditional Expectation

Let X, Y be jointly distributed random variables.
If X is discrete (and Y is either discrete or continuous), then:

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

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. These formulas are exactly the same as $\mathbb{E}\left[g(X)\right]$, except the PMF/PDF of X is replaced with the conditional PMF/PDF of $X \mid Y$.

Law of Total Expectation

Definition 5.3.72: Law of Total Expectation

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

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

This looks exactly like the law of total probability we are used to. Basically to solve for $\mathbb{E}\left[g(X)\right]$, we need to take a weighted average of $\mathbb{E}\left[g(X) \mid Y = y\right]$ over all possible values of y.

Covariance and Properties

Definition 5.4.73: Covariance

Let X, Y be random variables. The covariance of X and Y is:

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

Note: covariance can be negative, unlike variance.
Covariance satisfies the following properties:

1. If $X \perp Y$, then $Cov(X, Y) = 0$ (bot not necessarily vice versa, because the covariance could be zero but X and Y could not be independent).
2. $Cov(X, X) = Var(X)$.

3. $Cov(X, Y) = Cov(Y, X)$.

4. For scalars a, b, $Cov(aX + bY, Z) = a \cdot Cov(X, Z) + b \cdot Cov(Y, Z)$. This can be easily remembered like the distributive property of of sums $(aX + bY)Z = a(XZ) + b(YZ)$.

5. $Var(X + Y) = Var(X) + Var(Y) + 2 \cdot Cov(X, Y)$, and hence if $X \perp Y$, then $Var(X + Y) = Var(X) + Var(Y)$ (as we discussed earlier).

6. $Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_i) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_i)$. That is covariance works like FOIL (first, outer, inner, last) for multilication of sums $((a + b + c)(d + e) = ad + ae + bd + be + cd + ce)$.

(Pearson) Correlation

Definition 5.4.74: (Pearson) Correlation

Let X, Y be random variables. The (Pearson) correlation of X and Y is:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

We can prove by the Cauchy-Schwarz inequality (to be discussed later), $-1 \leq \rho(X, Y) \leq 1$. 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 ρ is the same as that of a.
In linear regression ("line-fitting") you may have calculated some R^2, $0 \leq R^2 \leq 1$, and this is actually ρ^2, and measure how well a linear relationship exists between X and Y. R^2 is the percentage of variance in Y which can be explained by X.

Variance of Sums of Random Variables

Fact 5.4.7: Variance of Sums of RVs

$$\begin{aligned}
Var(\sum_{i=1}^{n} X_i) &= Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j) && \text{[by definition of covariance]} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j) && \text{[by FOIL]} \\
&= \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j) && \text{[by symmetry (see image below)]}
\end{aligned}$$