Introduction

De Morgan's Laws

1. $(A \cup B)^C = A^C \cap B^C$
2. $(A \cap B)^C = A^C \cup B^C$

The 3 axioms of Probability

1. For every event $E \subset S$, $P(E) \geq 0$
2. $P(S) = 1$
3. If $E_1, E_2, \ldots$ is a sequence of events such that $E_i \cap E_j = \phi \ \forall i \neq j$, then
$P \bigcup_{i=1}^{\infty} E_i = \sum_{i=1}^{\infty} P(E_i)$

Introductory Theorems (Proofs not included):

1. $P(E^C) = 1 - P(E)$
2. $P(\phi) = 0$
3. $P(E \cap F^C) = P(E) - P(E \cap F)$
4. $E \subset F \Rightarrow P(E) \leq P(F)$
5. $P(E \cup F) = P(E) + P(F) - P(E \cap F)$
6. If S is a finite sample space with N equally probable outcomes, and E is any event in S, then $P(E) = |E|/N$

Sausage rule:
Given a problem where we are asked to choose a sample of size n from a space with N objects where a are tagged, and we are asked to find the probability of the event where x of the sample will be tagged.
Define 2 "sausages", the first has as objects sets of x tagged objects that can be drawn from the a tagged objects, the second has as objects sets of $n - x$ untagged objects that can be drawn from the $N - a$ untagged objects. We wish to count the number of ways that we can draw objects from "sausage" 1 and "sausage" 2.
Therefore, this type of problem can be solved with:

$$\frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}}$$

This later came to be known as a Hypergeometric Distribution.

Conditional Probability

Definition:
$P(B|A) = P(B \cap A)/P(A)$
By extension, we now have that
$P(B \cap A) = P(B|A)P(A) = P(A|B)P(B)$

Conditional Probability Theorems (Proofs not included):

1. $P(B|A) \geq 0 \ \forall B$
2. $P(S|A) = 1$
3. If $B_1,\ B_2,\ \ldots$ are disjoint events, then $P(\bigcup_{i=1}^{\infty} B_i|A) = \sum_{i=1}^{\infty} P(B_i|A)$
4. $P(B^C|A) = 1 - P(B|A)$
5. A is any event. Define $B_1,\ B_2,\ \ldots,\ B_n$ such that
   a. $B_i \cap B_j = \phi \ \forall \ i \neq j$
   b. $\bigcup_{i=1}^{n} B_i = S$

   Then, we have that $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$
6. A is any event. Define $B_1,\ B_2,\ \ldots,\ B_n$ such that
   a. $B_i \cap B_j = \phi \ \forall \ i \neq j$
   b. $\bigcup_{i=1}^{n} B_i = S$

   Then, we have that $P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} \ \forall \ k = 1,\ 2,\ \ldots,\ n$

   This is known as Bayes' Theorem.

Definition:
1. Two events A and B are **independent** if $P(B|A) = P(B)$ and/or $P(A|B) = P(A)$.
2. We may also say that these two events are **independent** if $P(A \cap B) = P(A)P(B)$.
3. By extension, we may say that the events $A_1,\ A_2,\ \ldots,\ A_n$ are **mutually independent** if $P(\bigcap_{j=1}^{k} A_{ij}) = \prod_{j=1}^{k} P(A_{ij}) \ \forall \mathit{subsets}\ A_{i1},\ A_{i2},\ \ldots,\ A_{ik}\ \mathit{of}\ A_1,\ A_2,\ \ldots,\ A_n \ \forall k = 1,\ 2,\ \ldots,\ n$.
4. Given the same definition for subsets as seen above, we say that $A_1,\ A_2,\ \ldots,\ A_n$ are **pairwise independent** if $P(A_{ij} \cap A_{ik}) = P(A_{ij})P(A_{ik}) \ \forall A_{ij},\ A_{ik}$.
5. An infinite collection of events $A_1,\ A_2,\ \ldots$ is **independent** iff every finite collection of $A_i s$ is independent according to Definition 3.

Related Theorems (Proofs not included):
1. If A and B are independent, then A^C and B^C are independent.
2. If A and B are disjoint, then they are independent iff either $P(A) = 0$ or $P(B) = 0$.

Random Variables

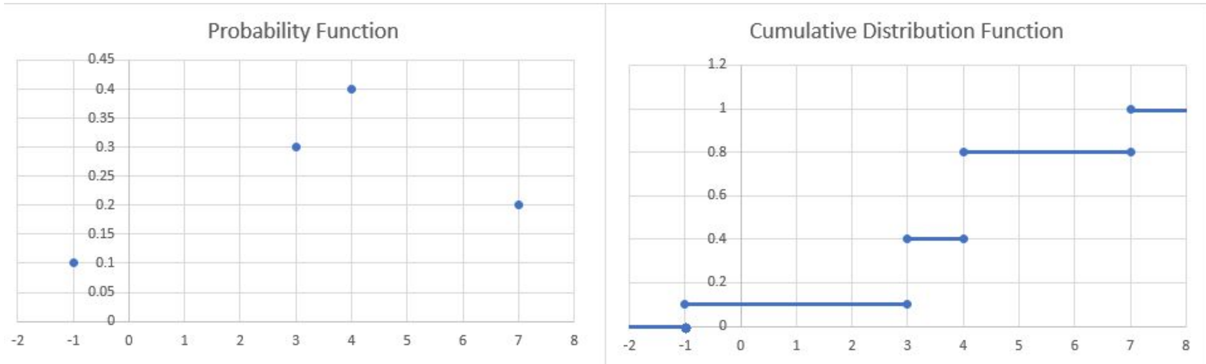
Definition:
Let S be a sample space with outcomes w. A **random variable** is defined to be a real-valued function $X : S \rightarrow \Re \ s.t.\ X(w) \varepsilon \Re \ \forall w \varepsilon S$.
The function of x given by $P(X \varepsilon]-\infty, x]) = P(X \leq x)$ is the **cumulative distribution function of X (cdfs)**, it may be denoted as $F_X(x)$. cdfs have the following properties:
1. $F_X(x)$ specifies probability distribution;
2. Define $P(X \varepsilon A) = P(w : X(w) \varepsilon A)$;

3. $F_X(x)$ has a few basic properties:
    a. $F_X(x) \le F_X(y)$ if $x \le y$;
    b. $F_X(\infty) = \lim_{x \to \infty} F_X(x) = 1$ and $F_X(-\infty) = \lim_{x \to -\infty} F_X(x) = 0$;
    c. $F_X(x)$ need only be continuous from the right.

<u>Discrete Random Variables</u>

Definition:
The real-valued function of x that gives $P(X = x)\ \forall x$ in the range of X is called the **probability function of X**, denoted by $p_X(x) = P(X = x)$.

Illustrating the difference between the cumulative distribution function of X and the probability function of X:



On the probability function, we see that $p_X(x_3) = P(X = x_3) = 0.4$
On the cumulative distribution function, we see that $F_X(x_3) = P(X\varepsilon] - \infty, x_3]) = P(X \le x_3) = 0.8$

<u>Named Discrete Random Variable Distributions</u>

1. Discrete Uniform Distribution
   $p_X(a_i) = P(X = a_i) = \frac{1}{N}\ \forall 1 \le i \le N$
   Used to model complete randomness over a discrete set.
2. Bernoulli Distribution
   $p_X(x) = P(X = x) = p^x(1-p)^{1-x}\ for\ 0 \le p \le 1\ and\ x = 0, 1$
   Used to describe random variables that take on one of two values.
3. Binomial Distribution
   $p_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}\ for\ 0 \le p \le 1,\ x = 0, 1, 2, ..., n$
   $X \sim Bin(n, p)$
   Arises in the “Binomial Setup”:
    a. We have n independent trials which each result in exactly one of two outcomes;
    b. The probability of success on trial i is consistently p.
4. Poisson Distribution
   $p_X(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}\ for\ x = 0, 1, 2, ...$

This is an approximation to the Binomial Distribution. To be used when, in the Binomial definition, n is large and p is small, we define $\lambda = np$.

5. Hypergeometric Distribution

   $$p_X(x) = P(X = x) = \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} \; for\ x = 0, 1, ..., min\{a, n\},\ a \le N,\ n \le N$$

   Arises when sampling without replacement from a set of N objects: a are of type 1, $N - a$ are of type 2, our desired sample has size n, and we want the probability of observing x of type 1 in our sample.

6. Geometric Distribution

   $p_X(x) = P(X = x) = (1 - p)^{x-1}p \; for\ x = 1, 2, ...$

   Arises when, given a sequence of independent and consistent Bernoulli trials, X denotes the trial number at which the first success is observed.

7. Negative Binomial Distribution

   $p_X(x) = P(X = x) = \binom{x-1}{k-1}p^k(1 - p)^{x-k} \; for\ x = k, k + 1, ...$

   Arises when, given a sequence of independent and consistent Bernoulli trials, X denotes the trial number at which the k^{th} success is observed.

<u>Mathematical Expectation and Variance of a Discrete Random Variable</u>

Definition:

The **Expected Value (or mean) of a discrete random variable X** is defined as

$$E(X) = \sum_{x\,:\,x \text{ is in the range of } X} x p_X(x) = \sum_{x\,:\,x \text{ is in the range of } X} xP(X = x)$$

$E(X) = \mu$

This is a weighted average of the values of X.

$E(X)$ has the following properties:

a) For some constant c, $E(cX) = cE(X)$
b) For some set of random variables $X_1, X_2, ..., X_n$, $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$

Definition:

The **Variance of a discrete random variable X** is defined as

$$Var(X) = {\sigma_X}^2 = E(X - \mu)^2 = \sum_{all\ x} (x - \mu)^2 P(X = x) = E(X^2) - \mu^2$$

The variance is the average of the squared distance of X from μ.

σ is the **standard deviation of a random variable**.

$E(X^2) \ge E(X)^2$ and $E(XY) \ne E(X)E(Y)$ (in general).

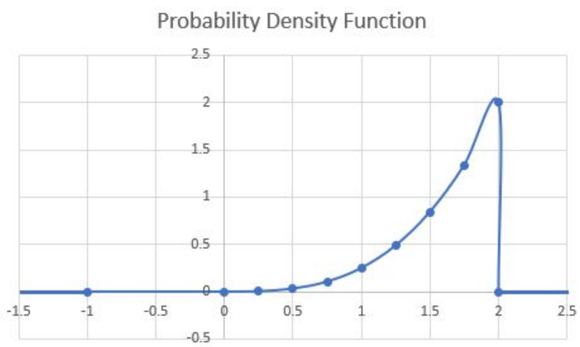
<u>Expectation and Variance of Some Above-Named Distributions:</u>

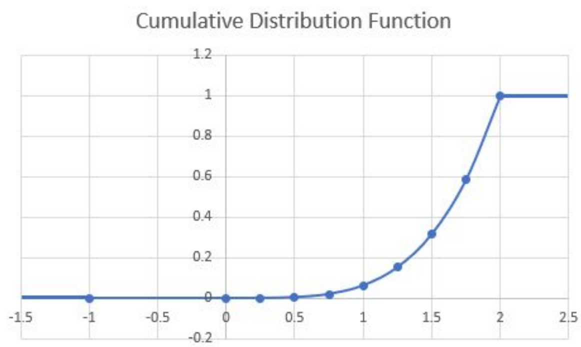
1. Discrete Uniform Distribution
   $E(X) = \sum_{i=1}^{N} a_i \frac{1}{N}$
   $Var(X) = \frac{1}{N} \sum_{i=1}^{N} (a_i - \sum_{j=1}^{N} (a_j \frac{1}{N}))^2 = \frac{1}{N} \sum_{i=1}^{N} (a_i - E(X))^2$
2. Bernoulli Distribution
   $E(X) = 1.p + 0.(1-p) = p$
   $Var(X) = E(X^2) - p^2 = p(1-p)$
3. Binomial Distribution
   $E(X) = np$
   $Var(X) = np(1-p)$
4. Poisson Distribution
   $E(X) = \lambda$
   $Var(X) = \lambda$
5. Geometric Distribution
   $E(X) = \frac{1}{p}$
   $Var(X) = \frac{1-p}{p^2}$

Useful property:
$E(X(X-1)) = E(X^2) - E(X)$
$Var(X) = E(X(X-1)) + E(X) - E(X)^2$

<u>Continuous Random Variables:</u>

Definition:
A random variable X is said to be **continuous** if its **cumulative distribution function** $F_X(x) = P(X \leq x)$ is a continuous function over all real values of x. As a consequence, $P(X = x) = 0$.
The **probability density function** $f_X(x)$ of a random variable X is a function with the following property: $P(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(y)dy$. Note that $\int_{-\infty}^{\infty} f_X(y)dy = 1$
Illustrating the difference between **cumulative distribution function** and **probability density function**:





Where $f_X(x) = \frac{x^3}{4}$ $when\ 0 \leq x \leq 2,\ f_X(x) = 0\ otherwise$.

Note: if we are given a real function $g(x) \geq 0$, we may convert it a probability density function $f_X(x)$ with $f_X(x) = \frac{g(x)}{\int_{-\infty}^{\infty} g(x)dx}$

<u>Mathematical Expectation and Variance of a Continuous Random Variable</u>

Definition:

The **Expected Value (or mean) of a continuous random variable X** is defined as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx$$

$$E(X) = \mu$$

Definition:

The **Variance of a continuous random variable X** is defined as

$$Var(X) = \sigma_X{}^2 = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)dx = E(X^2) - \mu^2 = \int_{-\infty}^{\infty} x^2 f_X(x)dx - \int_{-\infty}^{\infty} x f_X(x)dx$$

<u>Named Continuous Random Variable Distributions:</u>

1. The Uniform Distribution, $U(a,\ b)$
   $f_X(x) = \frac{1}{b-a}\ when\ a < x < b,\ 0\ elsewhere$
   $E(X) = \frac{a+b}{2}$
   $Var(X) = \frac{(b-a)^2}{12}$
   Arises in modelling a complete randomness in a continuous setting.
2. The Exponential Distribution, $exp(\beta)$
   $f_X(x) = \frac{1}{\beta}e^{\frac{-x}{\beta}}\ for\ x \geq 0,\ 0\ elsewhere$
   $E(X) = \beta$
   $Var(X) = \beta^2$
   Has the **memoryless property**, that is, $P(a < X < a+h \mid X > a) = P(0 < X < h)$
   The **hazard function** of X $\lambda_X(x) = \lim_{\Delta x \to 0} \frac{P(x<X<x+\Delta x \mid X>x)}{\Delta x}$. This function represents the ratio of the probability density function $f_X(x)$ to the survival function $P(X > x)$.
3. The Gamma Distribution, $\Gamma(\alpha,\ \beta)$
   $f_X(x) = \frac{1}{\Gamma(\alpha)}\frac{1}{\beta^\alpha}x^{\alpha-1}e^{\frac{-x}{\beta}},\ where\ \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}dx$
   $E(X) = \beta\alpha$
   $Var(X) = \beta^2\alpha$
   Arises in in processes for which the waiting times between Poisson-distributed events are relevant **[second opinion on this desired]**
4. The Chi-Squared Distribution, $\chi_\nu{}^2$
   $\chi_\nu{}^2 = \Gamma(\frac{\nu}{2},\ 2)$
   $E(X) = \nu$
5. The Normal Distribution, $N(\mu,\ \sigma^2)$
   $f_X(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{(\frac{-1}{2}\sigma^2(x-\mu)^2)}$
   $E(X) = \mu$
   $Var(X) = \sigma^2$

Arises in models extending the **central limit theorem**, which states that the sum of a large number of random variables is approximately normally distributed.
If $X \sim N(\mu,\ \sigma^2)$, we can define $Z = \frac{X-\mu}{\sigma}$ to get $Z \sim N(0,\ 1)$, this process is called **standardization**
This is analogous to the discrete Binomial Distribution

Transformations of Random Variables

Application:
We have the distribution of some Random Variable X and we wish to obtain the distribution of some real function $g(x)$ of X.

Transformations of Discrete Random Variables

1. Given a discrete random variable X, we wish to find the distribution of a discrete random variable Y which can be mapped by some real function $g(x)$ of X.
2. Find $P(Y = y) = P(g(x) = y) = P(X = g^{-1}(y))$
3. Substitute x with $g^{-1}(y)$ in the probability function of X, $p_X(x)$. The result is the probability function of Y, $p_Y(y)$.
4. Apply the same transformation to the domain of X to obtain the domain of Y.

Transformations of Continuous Random Variables

1. Given a continuous random variable X with pdf $f_X(x)$, we wish to find the pdf of a continuous random variable Y $f_Y(y)$, where Y is subject to an increasing function of X, $g(X)$
2. Find $F_Y(y) = F_X(g^{-1}(y))$
3. Find $f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}(g^{-1}(y))$

Definition:
Given some continuous random variable X which has a strictly increasing cdf $F_X(x)$, we can define a new random variable $Y = F_X(x)$, then $Y \sim U(0, 1)$. This transformation is called a **probability integral transformation**.

Theorem:
Given a random variable X and a random variable Y which is subject to the constraint $Y = g(X)$ such that $E|g(X)| < \infty$:

- If X is discrete, $E(g(x)) = \sum_{all\ x} g(x)p_X(x)$
- If X is continuous, $E(g(x)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx$

By the definition of $E(X)$.

Moment Generating Functions

Definition:

$E(X^k)$ is the k **th moment** of some random variable X if $E(|X|^k) < \infty$
For some random variable X, the function $M_X(t) = E(e^{tX})$ is called the **moment generating function of** X.

- If X is discrete, $M_X(t) = \sum_{all\ x} e^{tx} p_X(x)$
- If X is continuous, $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx$

Useful property: $M_X{}^{(k)}(0) = E(X^k)$

<u>Moment Generating Functions of some known distributions:</u>

- $X \sim Bin(n,\ p),\ M_X(t) = (e^t p + (1-p))^n$
- $X \sim Po(\lambda),\ M_X(t) = e^{\lambda(e^t - 1)}$
- $X \sim \Gamma(\alpha, \beta),\ M_X(t) = \frac{1}{(1-\beta t)^{\alpha}}$
- $X \sim N(\mu, \sigma^2),\ M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$
- $X \sim N(0, 1),\ M_X(t) = e^{\frac{t^2}{2}}$

Multivariate Distribution

Definition:
Given 2 random variables X and Y, $F_{X,Y}(x, y)$ is the **joint cumulative distribution function** of X and Y if $F_{X,Y}(x, y) = P(X \leq x \cap Y \leq y)$
The following properties must hold in order for two random variables X and Y to have a joint pdf:

- $f_{X,Y}(x, y) \geq 0$
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$ if X and Y are continuous
- $\int_{-\infty}^{y_0} \int_{-\infty}^{x_0} f_{X,Y}(x, y)dxdy = F_{X,Y}(x_0, y_0)$ if X and Y are continuous
- $\sum_{y:y \leq y_0} \sum_{x:x \leq x_0} p_{X,Y}(x, y) = F_{X,Y}(x_0, y_0)$ if X and Y are discrete

Given a **joint cumulative distribution function**, $f_X(x)$ $[p_X(x)]$ and $f_Y(y)$ $[p_Y(y)]$ are said to be the **marginal probability distribution functions** of X and Y.

Given some function of x and y, we can find the expected value of the function as follows:

- $E(g(x, y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)dxdy$ if X and Y are continuous
- $E(g(x, y)) = \sum_{all\ y} \sum_{all\ x} g(x, y) p_{X,Y}(x, y)$ if X and Y are discrete

Conditional Distributions

Given a pair of jointly distributed random variables X and Y, we define the **conditional distribution** of the two as

- $p_{Y|X=x}(y|x) = P(Y = y|X = x) = \frac{P(X=x,\ Y=y)}{P(X=x)}$ if X and Y are discrete and $P(X = x) \neq 0$
- $f_{Y|X=x}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$ if X and Y are continuous and $f_X(x)$

We can then define the **conditional cumulative distribution function** as

- $F_{Y|X=x}(y|x) = P(Y = y|X = x) = \int_{-\infty}^{y} f_{Y|X=x}(y|x)dy$ if X and Y are continuous
- $F_{Y|X=x}(y_0|x) = \sum_{y:y\leq y_0} \frac{p_{X,Y}(x,y_0)}{p_x(x)}$ if X and Y are discrete

We may then the **conditional expected value** as follows:

- $E(Y|X = x) = \int_{-\infty}^{\infty} yf_{Y|X=x}(y|x)dy$ if X and Y are continuous
- $E(Y|X = x) = \sum_{all\ y} yp_{Y|X=x}(y|x)$ if X and Y are discrete

Independence of Random Variables

Definition:
The random variables $X_1,\ X_2,\ ...,\ X_n$ are **independent** if

- $f_{X_1,X_2,...,X_n}(x_1x_2, ..., x_n) = f_{X_1}(x_1)f_{X_2}(x_2)..f_{X_n}(x_n)$ if all are continuous
- $p_{X_1,X_2,...,X_n}(x_1x_2, ..., x_n) = p_{X_1}(x_1)p(x_2)...p_{X_n}(x_n)$ if all are discrete
- $F_{X_1,X_2,...,X_n}(x_1x_2, ..., x_n) = F_{X_1}(x_1)F_{X_2}(x_2)...F_{X_n}(x_n)$

Property of independence:
Given 2 random variables X and Y, independent over functions $g(X),\ h(Y)$, we have that $E(g(X)h(Y)) = E(g(x))E(h(y))$

Theorem:
Let $X_1,\ X_2,\ ...,\ X_n$ be independent random variables;
Let $S_n = \sum_{i=1}^{n} X_i$;
Then $M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t)$

Theorem:
Let X, Y be discrete random variables with joint probability function $p_{X,Y}(x,y)$. Then:

- $p_Y(y) = \sum_x p_{Y|X=x}(y|x)p_X(x)$
- $E(Y) = \sum_x E(Y|X = x)p_X(x)$

Central Limit Theorem

Let $X_1, X_2, ..., X_n$ be independent random variables with $E(X) = \mu$ and $Var(X) = \sigma^2$. Then $P\left(\frac{\sum_{i=1}^{n} X_i - E(\sum_{i=1}^{n} X_i)}{\sqrt{Var(\sum_{i=1}^{n} X_i)}} \le x\right) \to P(Z \le x)$, where $Z \sim N(0, 1)$. This is equal to $P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \le x\right) \to P(Z \le x)$.