Foundations.

Definition. *The **sample space** Ω for a random scenario is the set of all possible outcomes. An **event** is a subset of Ω. A **probability measure** is a function*

$$\mathbb{P} : \Big\{ \textit{all events} \Big\} \to [0,1]$$

obeying three axioms:

(i) $\mathbb{P}(A) \geq 0 \quad \forall A \subset \Omega$.

(ii) $\mathbb{P}(\Omega) = 1$.

*(iii) If $A_1 \cap A_2 = \emptyset$ (in which case we say A_1 and A_2 are **disjoint** or **mutually exclusive**), then we have that*

$$\mathbb{P}\Big(A_1 \cup A_2\Big) = \mathbb{P}(A_1) + \mathbb{P}(A_2).$$

Suppose $B \subset \Omega$ is such that $\mathbb{P}(B) > 0$, then we can talk about the probability of an event $A \subset \Omega$ occurring given that we know B has occurred. This is denoted $\mathbb{P}(A|B)$, and is called the **conditional probability** of A given that B has occurred. We define

$$\mathbb{P}\big(A|B\big) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Theorem (Law of Total Probability). *Suppose $A_1, A_2 \subset \Omega$ partition our sample space. That is, $A_1 \cap A_2 = \emptyset$, and $A_1 \cup A_2 = \Omega$. Then,*

$$\mathbb{P}(B) = \mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2).$$

The above theorem generalizes in the obvious way for partitions consisting of more than two sets.

Theorem (Bayes' Rule). *Suppose $A, B \subset \Omega$ where $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then,*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}.$$

Two events A, B are said to be **independent** if

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

A **random variable** is a function $X : \Omega \to \mathbb{R}$. To each possible outcome in Ω, X assigns a real value. Each random variable X has an associated **probability distribution**. This essentially means that we can compute the probability that X takes a particular value, by summing the probabilities of all those outcomes in Ω which X maps to that value.
We say that two random variables X_1 and X_2 are **independent** if

$$\mathbb{P}(X_1 \in A, X_2 \in B) = \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B)$$

for all $A, B \subset \mathbb{R}$. Intuitively, you should think of random variables as being independent if knowledge of the outcome of one random variable gives no information about the outcome of the other random variable.

A random variable which can take on any value in a given (possibly infinite) interval of real numbers is said to be **continuous**.

A random variable which only takes on isolated values (for example, only integer values) is said to be **discrete**.

Discrete Random Variables. To each discrete random variable X we associate a **probability mass function** (or pmf) $p_X : \mathbb{R} \to [0,1]$ defined as follows

$$p_X(x) = \mathbb{P}\Big(\omega \in \Omega : X(\omega) = x\Big) = \mathbb{P}(X = x).$$

This probability mass function must satisfy

(i) $p_X(x) \geq 0 \quad \forall x \in \mathbb{R}$.

(ii) $p_X(x) > 0$ for at most countably many $x \in \mathbb{R}$.

(iii) $\displaystyle\sum_{x \in \mathbb{R}} p_X(x) = 1$.

Suppose we define a random variable $Y := g(X)$ via some function $g : \mathbb{R} \to \mathbb{R}$. Then its associated pmf is

$$p_Y(y) = \sum_{\{x \in \mathbb{R} : g(x) = y\}} p_X(x)$$

The **expectation** (or expected value, or **mean**) of a discrete random variable X is defined as

$$\mathbb{E}(X) = \sum_{x \in \mathbb{R}} x p_X(x).$$

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

$$\mathrm{Var}(X) = \mathbb{E}\bigg(\Big(X - \mathbb{E}(X)\Big)^2\bigg).$$

PROPOSITION. *We have the following basic properties. Suppose $a, b, c \in \mathbb{R}$.*

(i) *If $X > c$, then $\mathbb{E}(X) > c$.*
(ii) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.
(iii) $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
(iv) $\mathrm{Var}(X) = \mathbb{E}\Big(X^2\Big) - \Big(\mathbb{E}(X)\Big)^2$.
(v) *If $Y = g(X)$, then $\mathbb{E}(Y) = \mathbb{E}(g(X)) = \sum_{x \in \mathbb{R}} g(x) p_X(x)$.*

THEOREM. *If two discrete random variables X_1 and X_2 are independent, then*

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

Moreover, if $X_1, \ldots, X_n$ are pairwise independent, then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).$$

Since the definition of variance involves squaring the distance from X to its mean, the units in which variance is measured differ from those of the underlying random variable X. It is often expedient to measure the spread of a random variable in the same units as the random variable itself. To this end we introduce the **standard deviation** of X, $\mathrm{sd}(X)$, as an alternative measure of spread

$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}.$$

The **cumulative distribution function** (or cdf) $F_X : \mathbb{R} \to [0, 1]$ of a discrete random variable X is defined as

$$F_X(x) = \mathbb{P}(X \leq x).$$

There are many, many different probability distributions. We will content ourselves with three important examples.

EXAMPLE. *The **Bernoulli distribution** arises when conducting a single experiment with probability of success p. Let X represent the number of successes from the experiment.*

$$\begin{aligned} p_X(x) &= p^x (1-p)^{1-x} \quad x = 0, 1. \\ \mathbb{E}(X) &= p \\ \mathrm{Var}(X) &= p(1-p) \end{aligned}$$

EXAMPLE. *The **binomial distribution** arises when conducting n identical independent experiments, each with probability of success p. Let X represent the total number of successes from the n experiments.*

$$\begin{aligned} p_X(x) &= \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \cdots, n. \\ \mathbb{E}(X) &= np \\ \mathrm{Var}(X) &= np(1-p) \end{aligned}$$

EXAMPLE. *The **discrete uniform distribution** arises when an integer is uniformly selected from some interval* $[a, b]$.

$$p_X(x) = \frac{1}{b-a+1} \quad x = a, \cdots, b.$$
$$\mathbb{E}(X) = \frac{a+b}{2}$$
$$\mathrm{Var}(X) = \frac{(b-a)(b-a+2)}{12}$$

The following theorem is useful to know when working with the binomial distribution.

THEOREM (BINOMIAL THEOREM). *Let $n \in \mathbb{N}$ and $a, b \in \mathbb{R}$. Then,*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

Continuous Random Variables. A random variable is said to be **continuous** if there exists a non-negative function $f_X : \mathbb{R} \to \mathbb{R}$ such that

$$\mathbb{P}\Big(\omega \in \Omega : a \le X(\omega) \le b\Big) = \int_a^b f_X(x)\,dx.$$

For the rest of this document we shall write expressions such as that on the LHS above as $\mathbb{P}(a \le X \le b)$, in order to simplify the notation. The function f_X is called the **probability density function** (or pdf) of X. For any continuous random variable X we have the following properties

(i) $\mathbb{P}(X = x) = 0 \ \ \forall x \in \mathbb{R}$. Hence, for any $a < b$,

$$\mathbb{P}(a \le X \le b) = \mathbb{P}(a < X \le b) = \mathbb{P}(a \le X < b) = \mathbb{P}(a < X < b).$$

(ii) $\displaystyle\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$

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

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

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

$$\mathrm{Var}(X) = \mathbb{E}\bigg(\Big(X - \mathbb{E}(X)\Big)^2\bigg).$$

PROPOSITION. *Similar to before, we have the following basic properties. Suppose $a, b, c \in \mathbb{R}$.*

(i) *If $X > c$, then $\mathbb{E}(X) > c$.*
(ii) $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.
(iii) $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.
(iv) $\mathrm{Var}(X) = \mathbb{E}\Big(X^2\Big) - \Big(\mathbb{E}(X)\Big)^2$.
(v) *If $Y = g(X)$, then* $\mathbb{E}(Y) = \mathbb{E}(g(X)) = \displaystyle\int_{-\infty}^{\infty} g(x) f_X(x)\,dx.$

THEOREM. *If two continuous random variables X_1 and X_2 are independent, then*

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2).$$

Moreover, if $X_1, \ldots, X_n$ are pairwise independent, then

$$\mathrm{Var}(X_1 + X_2 + \cdots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n).$$

Similar to the discrete case, we introduce the **standard deviation** of X, $\mathrm{sd}(X)$, as an alternative measure of spread

$$\mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}.$$

The **cumulative distribution function** (or cdf) $F_X : \mathbb{R} \to [0, 1]$ of a continuous random variable X is defined as

$$F_X(x) = \mathbb{P}(X \le x).$$

F_X is continuous and non-decreasing. Moreover, it satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt.$$

By the Fundamental Theorem of Calculus we have that $F_X'(x) = f_X(x)$ at every x such that $f_X(x)$ is continuous.

For $0 \leq q \leq 100$ the value of $\lambda \in \mathbb{R}$ such that

$$\int_{-\infty}^{\lambda} f_X(x)\,dx = q$$

is called the q^{th} **percentile** of the distribution of the random variable X. The **median** of a distribution is defined to be the 50^{th} percentile of the distribution.

EXAMPLE. *The **continuous uniform distribution** arises when a real number is uniformly selected from some interval $[a, b]$. Suppose a random variable X has the continuous uniform distribution.*

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a,b], \\ 0 & x \notin [a,b]. \end{cases}$$

$$\mathbb{E}(X) = \frac{a+b}{2}$$

$$\operatorname{Var}(X) = \frac{(b-a)^2}{12}$$

The Normal Distribution. The most important continuous probability distribution is known as the **normal** (or **Gaussian**) **distribution**. There is an entire family of such distributions each member of which is uniquely specified by two parameters: its mean and variance. Suppose a random variable X is normally distributed with mean μ and variance σ^2 (and thus standard deviation of $\sigma > 0$), then we have

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

$$\mathbb{E}(X) = \mu$$

$$\operatorname{Var}(X) = \sigma^2$$

The following estimates apply to any normally distributed random variable.

(i) approximately 68% of the values lie within one standard deviation of the mean.
(ii) approximately 95% of the values lie within two standard deviations of the mean.
(iii) approximately 99.7% of the values lie within three standard deviations of the mean.

Note that the normal distribution with $\mu = 0$ and $\sigma = 1$ is called the **standard normal distribution**, and a random variable having this distribution is usually denoted by Z. The pdf of the standard normal distribution is given by $f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. Note that if X is normally distributed with mean μ and standard deviation σ, then $Z = (X - \mu)/\sigma$ will have the standard normal distribution.

Suppose we have a binomial random variable X with parameters n and p. Empirical observation of the graph of the associated probability mass function suggests that for large n we can fit the probability density function of a suitable normal distribution to this graph. This is indeed the case, the reasons for which are deeper than we will develop here. The guiding principle is as follows: If n is sufficiently large, the binomial random variable X will be approximately normally distributed with a mean of $\mu = np$ and a standard deviation of $\sigma = \sqrt{np(1-p)}$. How large is "sufficiently large" depends on the particular values of the parameters and the desired level of accuracy. In general, the larger n is, the better the approximation will be.

THEOREM. *Suppose X_1 and X_2 are independent normal random variables and $a_1, a_2 \in \mathbb{R}$. Then $a_1X_1 + a_2X_2$ is a normal random variable the mean and variance of which can be determined as follows*

$$\mathbb{E}(a_1X_1 + a_2X_2) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2).$$

$$\operatorname{Var}(a_1X_1 + a_2X_2) = a_1^2 \operatorname{Var}(X_1) + a_2^2 \operatorname{Var}(X_2).$$