Throughout this course, we will model data using random variables. The goal of this lecture is to review relevant definitions and concepts concerning random variables and their distributions.

### 2.1 Random variables and distributions

A **discrete random variable** $X$ can take a finite, or at most countably infinite, number of possible real values. We often use discrete random variables to model categorical data (for example, the outcome of a dice roll) and count data (for example, how many students are enrolled in S&DS 242).

The distribution of a discrete random variable $X$ can be specified by its **probability mass function (PMF)**: For each possible value $x$ that $X$ can take,

$$f_X(x) = \mathbb{P}[X = x].$$

Then for any subset $A$ of values that $X$ can take,

$$\mathbb{P}[X \in A] = \sum_{x \in A} f_X(x).$$

If $\mathcal{X}$ is the entire space of possible values for $X$, then $\sum_{x \in X} f_X(x) = \mathbb{P}[X \in \mathcal{X}] = 1$.

A **continuous random variable** $X$ can take any real value. We use continuous random variables to model continuous data (for example, the height or weight of a person). For any single value $x \in \mathbb{R}$, the probability that $X$ is exactly equal to $x$ is zero: $\mathbb{P}[X = x] = 0$. Instead, the distribution of $X$ may be specified by its **probability density function (PDF)** $f_X(x)$. This specifies that for any subset $A \subseteq \mathbb{R}$,

$$\mathbb{P}[X \in A] = \int_A f_X(x) \, dx.$$

The integral over the entire real line satisfies $\int_{-\infty}^{\infty} f_X(x) \, dx = \mathbb{P}[X \in \mathbb{R}] = 1$.

In both cases, when it is clear which random variable is being referred to, we will often simply write $f(x)$ for $f_X(x)$.

The distribution of $X$ can equivalently be specified by its **cumulative distribution function (CDF)**

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

For discrete random variables, this is given by

$$F_X(x) = \sum_{y : y \leq x} f_X(y).$$
This function $F_X(x)$ is a “staircase” or “step” function, with a jump of size $f_X(x) = \mathbb{P}[X = x]$ at each possible value $x$ for $X$. For continuous random variables, the CDF is given by

$$F_X(x) = \int_{-\infty}^{x} f_X(y) dy.$$  

This function $F_X(x)$ is continuous in $x$, and the fundamental theorem of calculus implies

$$f_X(x) = \frac{d}{dx} F_X(x).$$

By definition, $F_X$ is always increasing:

$$F_X(x) \leq F_X(y) \text{ if } x \leq y.$$  

We always have $F_X(x) = \mathbb{P}[X \leq x] \rightarrow 0$ as $x \rightarrow -\infty$, and $F_X(x) = \mathbb{P}[X \leq x] \rightarrow 1$ as $x \rightarrow \infty$. If $F_X$ is continuous and strictly increasing, meaning

$$F_X(x) < F_X(y) \text{ for all real numbers } x < y,$$

then $F_X$ has a continuous inverse function $F_X^{-1} : (0,1) \rightarrow \mathbb{R}$, which satisfies

$$F_X(F_X^{-1}(t)) = t \quad \text{and} \quad F_X^{-1}(F_X(x)) = x.$$  

This inverse function $F_X^{-1}$ is called the quantile function: For any value $t \in (0,1)$, the number $F_X^{-1}(t)$ is the value of $x$ for which $F_X(x) = \mathbb{P}[X \leq x] = t$. This value $x$ is called the $t^{th}$ quantile of the distribution of $X$. For example, when $t = 0.5$, $F_X^{-1}(0.5)$ is the median of the distribution of $X$. For $t = 0.25$ and $t = 0.75$, $F_X^{-1}(0.25)$ and $F_X^{-1}(0.75)$ are the first and third quartiles of the distribution of $X$.

## 2.2 Expected value and variance

For any random variable $X$, the expectation or mean of $X$ is its “average value”. If $X$ is discrete with PMF $f_X(x)$, then

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot f_X(x)$$

where the sum is over all possible values of $X$. If $X$ is continuous with PDF $f_X(x)$, then analogously,

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

More generally, for any function $g : \mathbb{R} \rightarrow \mathbb{R}$, the mean of $g(X)$ is

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) \cdot f_X(x)$$

2-2
for discrete $X$ and
\[
\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) \, dx
\]
for continuous $X$.

A very important property of the expectation is that it is linear: For any random variables $X_1, \ldots, X_n$ (not necessarily independent),
\[
\mathbb{E}[X_1 + \ldots + X_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n],
\]
and for any constant $c \in \mathbb{R}$,
\[
\mathbb{E}[cX] = c \mathbb{E}[X].
\]

The variance of $X$ is defined by the two equivalent expressions
\[
\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2] - (\mathbb{E}X)^2.
\]

It is invariant to translations: For any constant $c \in \mathbb{R}$, $\text{Var}[X + c] = \text{Var}[X]$. If $X$ is centered such that $\mathbb{E}X = 0$, then $\text{Var}[X] = \mathbb{E}[X^2]$.

For any constant $c \in \mathbb{R}$, $\text{Var}[cX] = c^2 \text{Var}[X]$. If $X_1, \ldots, X_n$ are independent (or more generally, pairwise uncorrelated), then
\[
\text{Var}[X_1 + \ldots + X_n] = \text{Var}[X_1] + \ldots + \text{Var}[X_n], \tag{2.1}
\]
If $X_1, \ldots, X_n$ are correlated, then this is not true—for example, we have
\[
\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2 \text{Cov}[X, Y]
\]
where $\text{Cov}[X, Y]$ is the covariance between $X$ and $Y$. (See Example 2.6 below).

The standard deviation of $X$ is $\sqrt{\text{Var}[X]}$. This has the benefit of being interpretable on the same scale as $X$, whereas $\text{Var}[X]$ is on the scale of $X^2$.

### 2.3 A few examples

**Example 2.1.** A Bernoulli random variable $X \sim \text{Bernoulli}(p)$ (for $p \in [0, 1]$) is a discrete random variable taking two possible values $\{0, 1\}$. Its PMF is given by
\[
f(x) = \begin{cases} 
p & \text{if } x = 1 \\
1-p & \text{if } x = 0. \end{cases}
\]
The mean of $X$ is
\[
\mathbb{E}[X] = 0 \cdot \mathbb{P}[X = 0] + 1 \cdot \mathbb{P}[X = 1] = p.
\]
The variance of $X$ is
\[
\text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = \mathbb{E}[X] - (\mathbb{E}X)^2 = p - p^2 = p(1-p),
\]
where we have used $X^2 = X$ in the second equality because $X \in \{0, 1\}$. 

2-3
Example 2.2. A Binomial random variable $X \sim \text{Binomial}(n, p)$ (for a positive integer $n$ and $p \in [0, 1]$) is a discrete random variable taking values in $\{0, 1, 2, \ldots, n\}$. Its PMF is given by

$$f(x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n-x} & \text{if } x \in \{0, 1, 2, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}$$

If $X_1, \ldots, X_n$ are independent Bernoulli($p$) random variables, then their sum $X = X_1 + \ldots + X_n$ is Binomial($n, p$)—hence this represents “the total number of heads in $n$ tosses of a coin that lands heads with probability $p$”.

Using the above properties of mean and variance, we have

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n] = np, \quad \text{Var}[X] = \text{Var}[X_1] + \ldots + \text{Var}[X_n] = np(1 - p).$$

(It is possible, but more difficult, to derive these also from the definitions of $\mathbb{E}[X]$ and $\text{Var}[X]$ and the above PMF.)

Example 2.3. A Normal random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ (for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$) is a continuous random variable taking any real value. Its PDF is given by

$$f(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

whose graph is a bell-shaped curve centered at $\mu$ with width proportional to $\sigma$.

The mean $\mathbb{E}[X]$ is $\mu$, which may be verified by a calculus exercise:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

$$= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} (y + \mu) \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy \quad \text{(change of variables } y = x - \mu)$$

$$= \int_{-\infty}^{\infty} y \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy + \mu \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{y^2}{2\sigma^2}} dy.$$

The first integral is 0, because the integrand is an odd function of $y$, so the integral over $y \in (0, \infty)$ cancels the integral over $y \in (-\infty, 0)$. The second integral is 1 because it is the integral of the $\mathcal{N}(0, \sigma^2)$ PDF over the entire real line. Hence $\mathbb{E}[X] = \mu$.

For the variance,

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

$$= \int_{-\infty}^{\infty} (x - \mu)^2 \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

$$= \int_{-\infty}^{\infty} \sigma^2 y^2 \cdot \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \cdot \sigma dy \quad \text{(change of variables } y = (x - \mu)/\sigma)$$

$$= \sigma^2 \int_{-\infty}^{\infty} y^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy.$$
Integration by parts shows that this last integral is 1: Let \( u = y \) and \( v = -\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \). Then \( du = dy \) and \( dv = y \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \), so

\[
\int_{-\infty}^{\infty} y^2 \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = \int_{-\infty}^{\infty} u dv = uv \bigg|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} v du
\]

\[
= 0 + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = 1.
\]

So \( \text{Var}[X] = \sigma^2 \), and \( X \) has standard deviation \( \sigma \).

\[\Box\]

**Example 2.4.** A Gamma random variable \( X \sim \text{Gamma}(\alpha, \beta) \) (for \( \alpha, \beta > 0 \)) is a continuous random variable taking positive real values. Its PDF is given by

\[
f(x) = \begin{cases} 
\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x > 0 \\
0 & x \leq 0
\end{cases}
\]

In the above, \( \Gamma : (0, \infty) \to (0, \infty) \) is the Gamma function, defined by the integral

\[
\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx.
\]

You may think of the Gamma function as extending the factorial function to all positive real numbers: For positive integers \( n \), we have \( \Gamma(n) = (n-1)! \). A calculus exercise (omitted here for brevity) verifies that

\[
\mathbb{E}[X] = \frac{\alpha}{\beta}, \quad \text{Var}[X] = \frac{\alpha}{\beta^2}.
\]

\[\Box\]

### 2.4 Joint distributions, independence, and covariance

The joint distribution of \( k \) different random variables \( X_1, \ldots, X_k \) may be specified, in the discrete case, by the **joint PMF**

\[
f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) = \mathbb{P}[X_1 = x_1, \ldots, X_k = x_k].
\]

In the continuous case, it may be defined by the **joint PDF** \( f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \). This specifies that for any subset \( A \subseteq \mathbb{R}^k \),

\[
\mathbb{P}[(X_1, \ldots, X_k) \in A] = \int_A f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) dx_1 \ldots dx_k.
\]

When it is clear which random variables are being referred to, we will write \( f(x_1, \ldots, x_k) \) for \( f_{X_1,\ldots,X_k}(x_1, \ldots, x_k) \).
Example 2.5. \((X_1, \ldots, X_k)\) have a multinomial distribution,

\((X_1, \ldots, X_k) \sim \text{Multinomial} \left( n, (p_1, \ldots, p_k) \right) \),

if these random variables take nonnegative integer values summing to \(n\), with joint PMF

\[
f(x_1, \ldots, x_k) = \begin{cases} 
\binom{n}{x_1, \ldots, x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k} & \text{if } x_1, \ldots, x_k \geq 0 \text{ and } x_1 + \ldots + x_k = n \\
0 & \text{otherwise.}
\end{cases}
\]

Here, \(p_1, \ldots, p_k\) are values in \([0,1]\) that satisfy \(p_1 + \ldots + p_k = 1\) (representing the probabilities of \(k\) different mutually exclusive outcomes), and \(\binom{n}{x_1, \ldots, x_k}\) is the multinomial coefficient

\[
\binom{n}{x_1, \ldots, x_k} = \frac{n!}{x_1! x_2! \cdots x_k!}.
\]

The multinomial distribution generalizes the binomial distribution to \(k > 2\) outcomes. It describes the number of samples belonging to each of the \(k\) outcomes, if there are \(n\) total samples, each independently belonging to outcomes 1, \ldots, \(k\) with probabilities \(p_1, \ldots, p_k\).

For example, if we roll a standard six-sided die 100 times and let \(X_1, \ldots, X_6\) denote the number of rolls that yielded 1 to 6, then

\((X_1, \ldots, X_6) \sim \text{Multinomial} \left( 100, \left( \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6} \right) \right) \).

Random variables \(X_1, \ldots, X_n\) are independent (in both the discrete and continuous settings) if

\[
f_{X_1, \ldots, X_n}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_n}(x_n).
\]

In particular, when \(X_1, \ldots, X_n\) are independent, their joint distribution is fully determined by the individual distributions of \(X_1, \ldots, X_n\).

If \(X_1, \ldots, X_n\) are independent, then this implies that for any subsets \(A_1, \ldots, A_n \subseteq \mathbb{R}\),

\[
P[X_1 \in A_1 \text{ and } \ldots \text{ and } X_n \in A_n] = P[X_1 \in A_1] \cdot \ldots \cdot P[X_n \in A_n].
\]

Furthermore, for any functions \(g_1, \ldots, g_n : \mathbb{R} \to \mathbb{R}\), we also have

\[
\mathbb{E}[g_1(X_1) \cdot \ldots \cdot g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdot \ldots \cdot \mathbb{E}[g_n(X_n)].
\]

The covariance between two random variables \(X\) and \(Y\) is defined by the two equivalent expressions

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

In particular, \(\text{Var}[X] = \text{Cov}[X, X]\). Like the variance, this is translation invariant: For any constants \(a, b \in \mathbb{R}\), \(\text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]\). If \(\mathbb{E}X = 0\) and \(\mathbb{E}Y = 0\), then \(\text{Cov}[X, Y] = \mathbb{E}[XY]\).

If \(X\) and \(Y\) are independent, then \(\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]\), so \(\text{Cov}[X, Y] = 0\). However, the converse is not true—Problem 2 of Homework 1 provides a counterexample.
An important property of the covariance is that it is \textit{bilinear}: For any random variables \(X_1,\ldots,X_n\) and \(Y_1,\ldots,Y_m\),

\[
\text{Cov}[X_1 + \ldots + X_n, Y_1 + \ldots + Y_m] = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}[X_i, Y_j],
\]

and for any constants \(a, b \in \mathbb{R}\),

\[
\text{Cov}[aX, bY] = ab \text{Cov}[X, Y].
\]

\textbf{Example 2.6.} This allows us to deduce a general expression for \(\text{Var}[X_1 + \ldots + X_n]\) in the setting where \(X_1,\ldots,X_n\) may be \textit{dependent}:

\[
\text{Var}[X_1 + \ldots + X_n] = \text{Cov}[X_1 + \ldots + X_n, X_1 + \ldots + X_n]
= \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}[X_i, X_j]
= \sum_{i=1}^{n} \text{Cov}[X_i, X_i] + 2 \sum_{i<j} \text{Cov}[X_i, X_j].
\]

The last sum above is over all \(\binom{n}{2}\) pairs of indices \(i, j \in \{1,\ldots,n\}\) where \(i < j\). Identifying \(\text{Var}[X_i] = \text{Cov}[X_i, X_i]\) in this last expression, we arrive at

\[
\text{Var}[X_1 + \ldots + X_n] = \sum_{i=1}^{n} \text{Var}[X_i] + 2 \sum_{i<j} \text{Cov}[X_i, X_j].
\]

If \(X_1,\ldots,X_n\) are independent, then \(\text{Cov}[X_i, X_j] = 0\) for all \(i < j\), so this becomes Eq. \((2.1)\).

The \textbf{correlation} between \(X\) and \(Y\) is their covariance normalized by the product of their standard deviations:

\[
\text{corr}(X, Y) = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}.
\]

Unlike the covariance, the correlation is dimension-free and invariant to rescaling: For any constants \(a, b > 0\), we have \(\text{corr}(aX, bY) = \text{corr}(X, Y)\). For any random variables \(X\) and \(Y\), we always have \(-1 \leq \text{corr}(X, Y) \leq 1\): This is a consequence of the \textbf{Cauchy-Schwarz inequality}

\[
\text{Cov}[X, Y]^2 \leq \text{Var}[X] \text{Var}[Y].
\]