3.1 Review

1. Announcements
   (1) HW1 due today
   (2) HW2 has been posted

2. Saw several "rules" to calculate probabilities:
   - Addition rule:
     \[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
   - Multiplication (chain) rule:
     \[ P(A \cap B) = P(A) P(B|A) \]
   - Law of total probability:
     \[ P(A) = P(A|B) P(B) + P(A|B^c) P(B^c) \]
   - Bayes' rule
     \[ P(A|B) = \frac{P(B|A) P(A)}{P(B)} \]
   - Many examples

3. Independence

\[ P(A_{i_1} \cap A_{i_2} \ldots \cap A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j}) \]

"Can multiply probabilities."
3.2 Independence again

Example 1: 3 out of 45 games that Arlene bowls, she scores at least 280 points. What are her chances of bowling a game over 280 points in her next 18 games?

\[ P(\text{Arlene scores } \leq 280 \text{ in 1 game}) = \left( \frac{14}{15} \right) \]

\[ P(\text{Arlene scores } \leq 280 \text{ in all 18 games}) = \left( \frac{14}{15} \right)^{18} \]

\[ P(\text{Scores over 280 at least once}) = 1 - \left( \frac{14}{15} \right)^{18} = 0.711. \]

Note: \( \left( \frac{14}{15} \right) \approx 0.933 \)

\( \left( \frac{14}{15} \right)^{18} \approx 0.288 \)

**Key point:** Probabilities are usually < 1, and then multiplying many probabilities leads to very small numbers.

Systems with many components.
3.3 Outline for today

1. Random variables.
2. CDF, PDF, PMF.
4. Joint distributions.
5. (Maybe) Multivariate distributions.

3.4 Random Variables

"Standard" definition: A random variable is a map $X : S \mapsto \mathbb{R}$, $S$ is the sample space of outcomes, and $\mathbb{R}$ is the real numbers.

Intuitively, sort of a lens through which we can "examine" the outcome, i.e. a measurement of a stochastic system.

Example 2: Flip a coin twice: $S = \{HH, HT, TH, TT\}$.

Consider, some random variables:

\[
\begin{align*}
X(HH) &= 2 & Y(HH) &= 1 \\
X(HT) &= 1 & Y(HT) &= 1 \\
X(TH) &= 1 & Y(TH) &= 0 \\
X(TT) &= 0 & Y(TT) &= 0.
\end{align*}
\]

$x = \# \text{ heads} \\
y = \frac{1}{\mathbb{1}_{\text{first toss landed heads}}}$.

**Define indicator.**
A note on the shorthand notation: Often,

\[ P(X = 1) = P(\exists \omega \in S: X(\omega) = 1) \]

Note: You can replace \( \mathbb{R} \) with any measurable space (think of things like \textit{word-valued} random variables), but then defining things like distribution functions etc. might get technical.

**Example 3: Indicator Random Variables:** Suppose that our experiment consists of measuring how long bulbs last in a household. Further assume we are not interested in precisely how long the bulbs last but just if they last over 5 years or not.

\[ I_b = \begin{cases} 
1 & \text{if bulb lasts over 5 years} \\
0 & \text{otherwise}
\end{cases} \]

\( I_b \) for each bulb.

Note: Indicators play nicely with expectations.
Example 4: A discrete random walk $W$, and its location at a particular time.

A random walk; on the integers

\[ \begin{array}{ccc}
-1 & 0 & 1 \\
\end{array} \]

$t=0$

After $k$ time steps, the position of the random walk is a random variable,

It takes values in the set $\{-k, -k+1, \ldots, 0, \ldots, k\}$

(i.e.) $X(W = w) \in \{-k, \ldots, k\}$. 
Example 5: Features in machine learning: The pixels of an image are random variables, we process these (by using filters of various sorts), and obtain features. Features are also random variables. Finally, we want to build a face-detector.

Functions of RVs are also RVs. So roughly, can think of face-detectors (more generally machine learning) as measuring some RVs (pixels) generating some intermediate RVs (features) and attempting to predict some other RVs (face or not).

Goal: "Predict" \( Y \leftrightarrow \) face or not.

Keypoint: Functions of RVs are also RVs.
3.5 Cumulative distribution function

A random variable is random. There is no way to determine its value before running the stochastic experiment. In order to understand random variables, and make a-priori decisions, we often consider non-random functions of the random variable (cumulative distribution functions, expectations, variances etc.).

The cumulative distribution function or CDF of a random variable $X$, is defined as,

$$F_X(x) = \Pr(X \leq x).$$

The CDF $F$ is defined for every $x$, and uniquely determines the distribution of the random variable $X$.

Useful as a way to characterize an RV, (ie) if I want to show that $X$ has a Gaussian distribution, I can calculate

$$F_X(x) \&\text{ compare to } \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{x^2}{2} \right\} \, dx.$$
Example 5: The Kolmogorov-Smirnov test:

In Statistics we often want to test if "samples" $X_1, \ldots, X_n$ have some fixed distribution (say Gaussian).

"Empirical" CDF

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}.$$

"True" CDF

$$f_z(x) = \mathbb{P}(Z \leq x)$$

Test Statistic:

$$T = \sup_x \left| F_n(x) - F_z(x) \right|$$

:-C if $T$ is large.
Example 6: Two fair coins are tossed and the outcome is observed. Before the coins are tossed, we are given the following choice of payoffs:

Payoff 1:
- Win $1 for each head.
- Lose $3 for getting two tails.

Payoff 2:
- Win $1 if the coins are different.
- Win $2 if both coins turn up tails.
- Lose $3 if both coins turn up heads.

Which payoff should we choose?

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>HH</td>
<td>2</td>
<td>-3</td>
</tr>
<tr>
<td>HT</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TH</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TT</td>
<td>-3</td>
<td>2</td>
</tr>
</tbody>
</table>

Important: X & Y have same distribution & same CDF. They are not equal.
3.6 Discrete and Continuous random variables

RVs are sometimes divided into two groups, discrete and continuous. There are RVs that are neither (sometimes called mixed).

Common discrete distributions: Bernoulli, Binomial, Poisson

Common continuous distributions: Gaussian, Exponential, ...

A discrete RV \( X \) typically has an associated probability mass function (pmf):

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

A continuous RV \( X \) has an associated probability density function (pdf), also denoted \( f_X(x) \), with the property that,

\[
\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) \, dx \quad \text{assuming } a \leq b.
\]

For a continuous random variable, we think of

\[
\mathbb{P}(X = x) = 0 \neq f_X(x)
\]
3.7 Expectations

Often, things we care about are properties of random variables: is it big or small? Did we win money or not? How much did we win?

Again, RVs are random and so not tangible before running the experiment.

The expectation is an important non-random property of a random variable that we can “calculate” before running the experiment.

For a discrete random variable,

\[ E(X) = \sum_{x} x f_{x}(x) \]

For a continuous random variable,

\[ E(X) = \int_{-\infty}^{\infty} x f_{x}(x) \, dx \]
Example 7: A point $X$ is chosen uniformly between $[0, 1]$. Let $T = X^2$. Find $E(T)$.

This is a continuous RV.

$$E(T) = \int_{0}^{1} x^2 1_{\{x \in [0, 1]\}} \, dx$$

$$= \int_{0}^{1} x^2 \, dx = \frac{1}{3}$$

Example 8: A bucket contains 100 balls, 30 of which are blue and the rest are red. We draw a random subset of 5 balls (uniformly). Let $B$ be the number of blue balls in the chosen subset, and define the random variable $I = \mathbb{1}_{B<3}$. Find the expectation, $E(I)$.

This is a discrete RV.

$$E(I) = \frac{1}{\binom{100}{5}} \sum_{\text{subsets}} \mathbb{1}_{\{\text{subset has 3 blue balls}\}}$$

$$= \frac{\binom{30}{3} \binom{70}{2}}{\binom{100}{5}} = \mathbb{P}(\text{subset has 3 blue balls})$$

Generally,

$$E(\mathbb{1}_{\{\text{event}\}}) = 1 \cdot \mathbb{P}(\text{event}) + 0 \cdot \mathbb{P}(\text{event}')$$
3.7.1 Expectations as averages

An intuitive way to think about an expectation is as the average value of a random variable. Roughly, if I measured a random variable many times (independently), and took their average, I would obtain the expectation (formally, this is the law of large numbers).

Averages follow certain intuitive properties, and expectations do as well.

1. The constancy rule: For any constant $c$.
   \[ E[c] = c. \]

2. The scaling rule: For any constant $c$
   \[ E[cX] = cE[X]. \]

3. The ordering rule: If $X \leq Y$, then
   \[ EX \leq EY. \]

4. The additivity rule: Linearity of expectations.
   \[ E[X + Y] = EX + EY. \]
3.8 Variance

For a random variable, the variance is a measure of the spread of its distribution.

Formally,

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

"On average how far is \( X \) from its mean."

3.9 Pairs of random variables

Given a pair of random variables we can measure various "cross" quantities, and this leads to various definitions.

3.9.1 Co-variance

The covariance between two random variables \( X \) and \( Y \) is defined as:

\[ \text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \]

RVs "co-vary" if they fluctuate around the mean together & conversely.
3.9.2 Properties of variance and co-variance

Like the expectation, there are some simple rules that help calculate the variance of transformations of a random variable:

1. \[ \text{V}(X) = E \left( (X - E(X))^2 \right) \]
   \[ = E \left[ X^2 + (E(X))^2 - 2X \cdot E(X) \right] = E(X^2) - (E(X))^2 \]

2. \[ \text{V}(X + Y) = \text{V}(X) + \text{V}(Y) + 2 \text{cov}(X,Y) \]
   No linearity, unless \( \text{cov}(X,Y) = 0 \).

3. \[ \text{V}(aX + b) = a^2 \text{V}(X) \]

4. \[ \text{V}(\text{constant}) = 0 \]
3.9.3 Correlation

The (Pearson) correlation is a standardized version of the covariance.

\[
\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)} \sqrt{\text{var}(Y)}}
\]

Random variables with 0 correlation are said to be uncorrelated.

Alternatively,

\[
\text{cov}(X, Y) = 0 \implies \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]
\]
\[
= \mathbb{E}[XY - \mathbb{E}X \mathbb{E}Y - X \mathbb{E}Y + \mathbb{E}X \mathbb{E}Y]
\]
\[
= \mathbb{E}XY - \mathbb{E}X \mathbb{E}Y = 0.
\]
\[
\implies \mathbb{E}(XY) = \mathbb{E}X \mathbb{E}Y.
\]

Uncorrelated \(\implies\) \(\text{no "linear" association.}\)
3.9.4 Independence

A much stronger condition is \textit{independence}. Independence means that you cannot predict one random variable using the other. Roughly, uncorrelated means you cannot predict the other random variable using a linear function. Formally, two random variables are independent if

\[ \mathbb{E}[f(X)g(Y)] = [\mathbb{E}f(X)][\mathbb{E}g(Y)]. \]

For continuous \textit{fns} \( f \) \& \( g \).

"Every transformation of \( XY \) are uncorrelated."

\[ \mathbb{E}[f(X)g(Y)] = [\mathbb{E}f(X)][\mathbb{E}g(Y)]. \]

For continuous \textit{fns} \( f \) \& \( g \).

"Every transformation of \( XY \) are uncorrelated."
Example 9: Consider, $X \sim N(0,1)$ and $Y = X^2$. Are these random variables uncorrelated? Are they independent?

Side note: $Y \sim X_i^2$ with one degree of freedom.

$$\text{cov}(x, y) = E[(x - EX)(y - EY)]$$

$EX = 0$, $EX^2 = 1 = EY$.

$$= E \left[ x(x^2 - 1) \right] = EX^3 - EX$$

$$= 0.$$

"Every odd moment of a symmetric distribution (about 0) is $O(\text{if it exists})."
3.10 Joint distributions

We use the notation \( f_{(X_1, X_2, \ldots, X_n)}(x_1, x_2, \ldots, x_n) \) for the joint pdf/pmf of random variables \( X_1, X_2, \ldots, X_n \).

Discrete case:

\[
f_{(X_1, X_2, \ldots, X_n)}(x_1, x_2, \ldots, x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)
\]

Continuous case:

\[
P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_n}^{b_n} f_{(X_1, \ldots, X_n)}(x_1, \ldots, x_n) \, dx_1 \cdots dx_n.
\]

Assume \( a_i \leq b_i \quad \forall i \).
Example 10: A nut company markets cans of mixed nuts containing almonds, cashews and peanuts. Let $X$ and $Y$ represent the proportion of almonds and cashews respectively. Suppose that the joint pdf of $(X,Y)$ is given by:

$$f_{X,Y}(x,y) = \begin{cases} 24xy & \text{if } x > 0, y > 0, x + y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

- What is the probability that the almonds and cashew are at most 50% of the can? 😊
- What is $P(X > Y)$?

**Quick exercise:**
Verify this is a valid pdf.

**Picture:**

$$x + y = 1$$
\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{x,y}(x,y) \, dx \, dy = \int_{0}^{1} \int_{0}^{1-x} 24xy \, dx \, dy \\
= \int_{0}^{1} 24x \left[ \frac{y^2}{2} \right]_{0}^{1-x} \, dx \\
= 12 \int_{0}^{1} x (1+x^2-2x) \, dx = 12 \left[ \frac{x^3}{2} + \frac{x^4}{4} - \frac{2x^3}{3} \right]_{0}^{1} \\
= 1.
\]

(a) \( P(X + Y \leq 0.5) = \int_{0}^{\frac{1}{2}} \int_{0}^{\frac{1}{2}} 24xy \, dx \, dy \\
= 12 \int_{0}^{\frac{1}{2}} x (\frac{1}{2} - x)^2 \, dx = 12 \int_{0}^{\frac{1}{2}} x \left( \frac{1}{4} + x^2 - x \right) \, dx \\
= 12 \left[ \frac{x^3}{8} + \frac{x^4}{4} - \frac{x^3}{3} \right]_{0}^{1/2} \\
= 12 \left[ \frac{1}{32} + \frac{1}{16} + \frac{1}{64} - \frac{1}{24} \right] = \frac{1}{16}.
\]

(b) \( Y \)

Verify that symmetric \&

\( P(X > Y) \)

so \( P(X > Y) = \frac{1}{2} \).
Example 11: Multivariate normal: Often in statistics we want to model the distribution of random vectors. A common model is one that generalizes a univariate normal.

Univariate normal:  \[ f_X(x) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}. \]

Multivariate normal:

\[ f_{x,y}(x,y) = \frac{1}{(\sqrt{2\pi})^2 |\Sigma|^{\frac{1}{2}}} \exp \left\{ -\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2} \right\}, \]
\[ \mu \in \mathbb{R}^2, \quad \Sigma \in \mathbb{R}^{2 \times 2} \]