APM 421 Probability Theory
Conditional Probabilities

Jay Taylor
Fall 2013
Conditional Probabilities

Suppose that we have specified a probability distribution \( \mathbb{P} \) on a space \( S \) and that we learn that an event \( B \subseteq S \) has occurred. How should we update our beliefs concerning the probabilities of other events \( A \subseteq S \)?

Example: Suppose that a coin is tossed.

1. Assign a probability to the event \( A \) that the coin will land on heads.
Conditional Probabilities

Suppose that we have specified a probability distribution $\mathbb{P}$ on a space $S$ and that we learn that an event $B \subset S$ has occurred. How should we update our beliefs concerning the probabilities of other events $A \subset S$?

**Example:** Suppose that a coin is tossed.

1. Assign a probability to the event $A$ that the coin will land on heads.
2. Now, suppose that you are told that the same coin was previously tossed 20 times in succession and that it landed on heads 18 times. Assign a probability to $A$ taking into account this new information.
Conditional Probabilities

Suppose that we have specified a probability distribution $\mathbb{P}$ on a space $S$ and that we learn that an event $B \subset S$ has occurred. How should we update our beliefs concerning the probabilities of other events $A \subset S$?

**Example:** Suppose that a coin is tossed.

1. Assign a probability to the event $A$ that the coin will land on heads.

2. Now, suppose that you are told that the same coin was previously tossed 20 times in succession and that it landed on heads 18 times. Assign a probability to $A$ taking into account this new information.

The question that we wish to address here is whether there is a systematic way to update the probability of $A$ when we learn that $B$ is true.
To answer this question, we will modify the betting game introduced in earlier lectures by allowing the players to place bets on conditional events. A conditional bet is one that results in an exchange of money if and only if the event that is being conditioned on actually occurs.

To be concrete, suppose that player two pays $Pr(A|B)$ to bet $1 on the conditional event that $A$ occurs given that $B$ has occurred, denoted $A|B$. Depending on the outcome, player two will incur the following costs and profits:

- If $B$ occurs and $A$ occurs, then player one keeps the fee $Pr(A|B)$ but pays $1 to player two.
- If $B$ occurs but $A$ does not occur, then player one keeps the fee $Pr(A|B)$ and pays nothing to player two.
- If $B$ does not occur, then player one returns the fee $Pr(A|B)$ to player two and neither player makes a profit. In this event, we say that the bet has been called off.
Suppose that the prices assigned by player one to the wagers $AB$, $A|B$, and $B$ satisfy the inequality

$$Pr(AB) < Pr(A|B)Pr(B).$$

I claim that player two can make a Dutch book against player one by betting $1$ on $AB$ and selling bets of $1$ on $A|B$ and $\Pr(A|B)$ on $B$. The following table summarizes the profits and losses incurred by player two:

| event | $AB$ | $A|B$ | $B$ | total |
|-------|------|------|-----|-------|
| $\bar{B}$ | $-Pr(AB)$ | 0 | $Pr(A|B)Pr(B)$ | $Pr(A|B)Pr(B) - Pr(AB)$ |
| $\bar{AB}$ | $-Pr(AB)$ | $Pr(A|B)$ | $Pr(A|B)Pr(B) - Pr(A|B)$ | $Pr(A|B)Pr(B) - Pr(AB)$ |
| $AB$ | $1 - Pr(AB)$ | $Pr(A|B) - 1$ | $Pr(A|B)Pr(B) - Pr(A|B)$ | $Pr(A|B)Pr(B) - Pr(AB)$ |

Provided that the prices set by player one satisfy the inequality given above, it is clear that player two will earn a profit no matter which outcome occurs and so player two has created a Dutch book.
Similarly, if $Pr(AB) > Pr(A|B)Pr(B)$, then player two can create a Dutch book by betting $1$ on both $A|B$ and $B$ and selling a bet of $1$ on $AB$. These examples show that unless the prices assigned by player one to the three events $AB$, $B$ and $A|B$ satisfy the identity

$$Pr(AB) = Pr(A|B)Pr(B),$$

player two can arrange it so that player one is certain to lose money.

In fact, the converse of this result is also true: if the prices assigned by player one do satisfy this identity, then player two cannot make a Dutch book against them. To prove this, suppose that player two places bets of $w(B)$, $w(AB)$ and $w(A|B)$ on these three events. Notice that the amount of money that player two earns from the conditional bet on the event $A|B$ is a random variable that can be written as

$$w(A|B) \cdot I_B(I_A - Pr(A|B)).$$

Indeed, if the event $B$ does not occur, then the indicator variable $I_B = 0$, which means that this quantity vanishes, reflecting the fact that the conditional bet has been called off.
Using the result from the previous slide, the total amount of money that player two will either earn or lose from these three bets is a random variable that can be written as:

\[ W = w(B) \cdot (I_B - Pr(B)) + w(AB) \cdot (I_{AB} - Pr(AB)) + w(A|B) \cdot (I_B(I_A - Pr(A|B))). \]

Recalling that the expectation of an indicator variable is equal to the probability of the event that it indicates, the expected values of each of the three terms in this expression are:

\[
\begin{align*}
E[w(B) \cdot (I_B - Pr(B))] & = w(B) \cdot (Pr(B) - Pr(B)) = 0; \\
E[w(AB) \cdot (I_{AB} - Pr(AB))] & = w(AB) \cdot (Pr(AB) - Pr(AB)) = 0; \\
E[w(A|B) \cdot (I_B(I_A - Pr(A|B)))]) & = w(A|B) \cdot E[I_{AB} - Pr(A|B) \cdot I_B] \\
& = w(A|B) \cdot (Pr(AB) - Pr(A|B)Pr(B)) \\
& = 0,
\end{align*}
\]

where the last identity holds because we are assuming that \( Pr(AB) = Pr(B)Pr(A|B) \).
Putting these results together, we have shown that player two’s expected winnings are equal to 0, i.e., $\mathbb{E}[W] = 0$. As was the case when we proved the coherence theorem, this implies that one of the two following conditions is true:

- Either $W$ is trivial, i.e., $\mathbb{P}(W = 0) = 1$, in which case neither player earns or loses any money;
- or there is a positive probability that player two will lose money; in particular, player two cannot make a Dutch book against player one.

These results can be summarized as follows: When betting on a conditional event $A|B$, a necessary and sufficient condition for player one to avoid being a sure loser is that

$$Pr(AB) = Pr(A|B)Pr(B).$$
This last result can be used to justify the following definition of conditional probabilities. This is one of the most important definitions that we will encounter in this course.

**Definition**

Let $\mathbb{P}$ be a probability distribution on a set $S$ and suppose that $A$ and $B$ are events in $S$ with $\mathbb{P}(B) > 0$. Then the **conditional probability** $\mathbb{P}(A|B)$ of $A$ given $B$ is defined to be

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

In other words, once we have assigned probabilities to all of the events in a sample space $S$, subjectively or otherwise, the conditional probabilities are fully determined and no additional room for subjectivity remains.
Example: Let $X$ be a random variable that specifies the number of colds that you will have this year and suppose that $X$ has the following distribution:

\[
\begin{align*}
P(X = 0) &= 0.1, \\
P(X = 1) &= 0.4, \\
P(X = 2) &= 0.3, \\
P(X = 3) &= 0.2.
\end{align*}
\]

Let $A$ be the event that you will have exactly two colds, $A = \{X = 2\}$, and let $B$ be the event that you will have at least one cold, $B = \{X \geq 1\}$. Find the conditional probabilities $P(A|B)$ and $P(B|A)$. 

Using the definition of conditional probabilities and the fact that $A \subset B$, we have

\[
P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.3}{0.9} = \frac{1}{3};
\]

\[
P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A)}{P(A)} = 1.
\]

Notice that $P(A|B) \neq P(B|A)$, i.e., conditional probabilities are not symmetric.
**Example:** Let $X$ be a random variable that specifies the number of colds that you will have this year and suppose that $X$ has the following distribution:

\[
P(X = 0) = 0.1, \quad P(X = 1) = 0.4, \quad P(X = 2) = 0.3, \quad P(X = 3) = 0.2.
\]

Let $A$ be the event that you will have exactly two colds, $A = \{X = 2\}$, and let $B$ be the event that you will have at least one cold, $B = \{X \geq 1\}$. Find the conditional probabilities $P(A|B)$ and $P(B|A)$.

Using the definition of conditional probabilities and the fact that $A \subset B$, we have

\[
P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.3}{0.9} = 1/3;
\]

\[
P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A)}{P(A)} = 1.
\]

Notice that $P(A|B) \neq P(B|A)$, i.e., conditional probabilities are **not** symmetric.
Notice that conditional probabilities are only defined when the event that we are conditioning on has a positive probability. For example, the expression $P(A|\emptyset)$ has no meaning because $P(\emptyset) = 0$. To understand why we have not extended the definition to this case, suppose that $A$ and $B$ are events such that $P(B) = 0$. Then, since $B$ is the disjoint union of the events $AB$ and $\overline{AB}$, it follows that

$$0 = P(B) = P(AB) + P(\overline{AB}),$$

which shows that $P(AB) = P(\overline{AB}) = 0$. From our investigation of conditional bets, we know that the prices assigned to the three events $B$, $AB$ and $A|B$ should satisfy the condition

$$Pr(AB) = Pr(A|B) \cdot Pr(B).$$

However, when $Pr(B) = Pr(AB) = 0$, this identity will be satisfied no matter what value is assigned to $Pr(A|B)$. In other words, there is no natural unique way to assign a conditional probability to the event $A$ when the event being conditioned on has probability zero.
The following result, which is known as the **product rule**, is an immediate but very useful consequence of our definition of conditional probabilities.

**Theorem**

*Suppose that $A$ and $B$ are events and that $\mathbb{P}(B) > 0$. Then*

\[
\mathbb{P}(AB) = \mathbb{P}(A|B) \cdot \mathbb{P}(B).
\]

In words, the probability that both $A$ and $B$ are true is equal to the product of the conditional probability that $A$ is true given $B$ and the probability that $B$ is true. When $\mathbb{P}(A) > 0$, we can also write the probability that both $A$ and $B$ are true as

\[
\mathbb{P}(AB) = \mathbb{P}(B|A)\mathbb{P}(A).
\]

Whether we should condition on $A$ or on $B$ depends on the details of the problem. Usually it will be true that only one of these two expressions leads to probabilities that are easier to calculate than $\mathbb{P}(AB)$ itself.
Example: Suppose that you sample two balls at random and without replacement from an urn that contains five red balls and five blue balls. (Sampling without replacement means that when you choose a ball, it is permanently removed from the urn.) Find the probability that the first ball sampled is red and that the second ball sampled is blue.
Example: Suppose that you sample two balls at random and without replacement from an urn that contains five red balls and five blue balls. (Sampling without replacement means that when you choose a ball, it is permanently removed from the urn.) Find the probability that the first ball sampled is red and that the second ball sampled is blue.

Solution: Let $R$ be the event that the first ball sampled is red and let $B$ be the event that the second ball sampled is blue. Since we are sampling the balls in a particular order, it makes sense to condition on the earlier of the two events, i.e., on $R$. Then, by the product rule,

$$\mathbb{P}(RB) = \mathbb{P}(R)\mathbb{P}(B|R)$$

$$= \frac{5}{10} \times \frac{5}{9}$$

$$= \frac{5}{18}.$$
We can also use the product rule to calculate the probabilities of intersections containing more than two events. For example, if $A$, $B$ and $C$ are events and $P(ABC) > 0$, then

$$P(ABC) = P(C|AB)P(B|A)P(A).$$

As with pairwise intersections, we can condition in any order, so the following expression is equally valid:

$$P(ABC) = P(B|AC)P(C|A)P(A).$$

In general, if $A_1, \ldots, A_n$ are events and $P(A_1A_2\cdots A_n) > 0$, then

$$P(A_1A_2\cdots A_n) = P(A_1|A_2\cdots A_n)P(A_2|A_3\cdots A_n)\cdots P(A_n).$$

In this case, there are $n!$ different ways of ordering the sets and each of these leads to a legitimate (although not necessarily very useful) product expression for the probability on the left-hand side.
The Birthday Problem

**Problem:** Suppose that there are $k$ people in a room and that each person’s birthday is equally likely to be any of the 365 days that occur in a non-leap year. Calculate the probability, $s_k$, that at least two people in the room share the same birthday.

We will solve this problem by calculating the probability $t_k = 1 - s_k$ that no two people in the room share the same birthday. Suppose that the $k$ individuals are randomly assigned numbers $1, \cdots, k$, and let $A_j$ be the event that the first $j$ individuals do not share any birthdays in common with each other. Since the numbers are assigned at random (e.g., we do not selectively assign smaller numbers to individuals born earlier in the year), it follows that

$$t_j = P(A_j).$$

In particular, notice that $t_1 = 1$ simply because in a set containing only one individual there isn’t a second individual to share their birthday in common with them.
Now, by the product rule, we know that

\[ t_k = \mathbb{P}(A_k) = \mathbb{P}(A_k | A_{k-1}) \mathbb{P}(A_{k-1}) = \mathbb{P}(A_k | A_{k-1}) t_{k-1}. \]

Although we don’t yet know the value of \( t_{k-1} \), I claim that we can easily calculate \( \mathbb{P}(A_k | A_{k-1}) \), which is just the conditional probability that the \( k \)'th individual in the group does not share a birthday in common with any of the first \( k - 1 \) people given that the first \( k - 1 \) people do not share a birthday in common with each other. However, if \( A_{k-1} \) is true, then these \( k - 1 \) people must have \( k - 1 \) different birthdays, in which case the probability that the \( k \)'th person does not share birthday in common with any of these is equal to

\[ \mathbb{P}(A_k | A_{k-1}) = \frac{365 - (k - 1)}{365} = \left( 1 - \frac{k - 1}{365} \right). \]

This shows that \( t_k \) and \( t_{k-1} \) are related by the equation:

\[ t_k = \left( 1 - \frac{k - 1}{365} \right) t_{k-1}. \]
Continuing in this fashion, we also know that

\[ t_{k-1} = \left( 1 - \frac{k - 2}{365} \right) t_{k-2}, \]

which, in turn, implies that

\[
\begin{align*}
    t_k &= \left( 1 - \frac{k - 1}{365} \right) \left( 1 - \frac{k - 2}{365} \right) t_{k-2} \\
    &= \left( 1 - \frac{k - 1}{365} \right) \left( 1 - \frac{k - 2}{365} \right) \left( 1 - \frac{k - 3}{365} \right) t_{k-3} \\
    &= \ldots \\
    &= \left( 1 - \frac{k - 1}{365} \right) \left( 1 - \frac{k - 2}{365} \right) \left( 1 - \frac{k - 3}{365} \right) \ldots \left( 1 - \frac{1}{365} \right) t_1 \\
    &= \prod_{i=1}^{k-1} \left( 1 - \frac{i}{365} \right),
\end{align*}
\]

since \( t_1 = 1 \), as shown on the preceding slide.
The graph shows the probabilities $s_k$ plotted as a function of $k$ ranging from 1 to 40. Notice that the smallest value of $k$ such that $s_k > 0.5$ is $k = 23$ ($s_{23} \approx 0.5073$).
Although we have found an exact expression for $s_k$ and $t_k$, the fact that this is a product involving $k$ terms makes it difficult to interpret. However, when $k \ll 365$, $t_k$ can be approximated by a very simple expression. To understand this approximation, recall that the Taylor series expansion for the exponential function is

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{1}{2}x^2 + \cdots.$$ 

In particular, when $x$ is small in magnitude,

$$e^x \approx 1 + x \quad \text{and} \quad e^{-x} \approx 1 - x.$$ 

It follows that for values of $k \ll 365$, the probability $t_k$ can be approximated by:

$$t_k = \prod_{i=1}^{k-1} \left(1 - \frac{i}{365}\right) \approx \prod_{i=1}^{k-1} e^{-i/365} = \exp \left\{ -\frac{1}{365} \sum_{i=1}^{k-1} i \right\} = \exp \left\{ -\frac{1}{365} \frac{k(k-1)}{2} \right\}.$$
To derive this last identity, we used the fact that

\[ \sum_{i=1}^{k-1} i = \binom{k}{2} = \frac{k(k - 1)}{2}, \]

which can be verified by induction or by counting the number of off-diagonal vertices in a square lattice on the integers \( \{1, \cdots, k\} \times \{1, \cdots, k\} \).

The approximation

\[ t_k \approx \exp \left\{ -\frac{1}{365} \frac{k(k - 1)}{2} \right\} \]

can be interpreted as follows. When \( k \) is much smaller than 365, there are \( \binom{k}{2} \) pairs of individuals in the group, each of which has probability \( 1 - \frac{1}{365} \) of having distinct birthdays. Although these are not independent of one another, as long as \( k \ll 365 \), the dependence will be weak and so the probability that every pair has a distinct birthday will be approximately equal to \( 1 - \frac{1}{365} \) raised to the number of pairs.
The probabilities of complex events can often be evaluated by conditioning on additional information and using the following result.

**Theorem**

*(Law of Total Probability)* Suppose that $F_1, \cdots, F_n$ are mutually exclusive events with $\mathbb{P}(F_i) > 0$ for $i = 1, \cdots, n$ and let $E$ be an event with $E \subset F_1 \cup \cdots \cup F_n$. Then

$$
\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \cap F_i) = \sum_{i=1}^{n} \mathbb{P}(E|F_i)\mathbb{P}(F_i).
$$

For this result to be useful, the probabilities $\mathbb{P}(F_i)$ and $\mathbb{P}(E|F_i)$ must be easier to calculate than $\mathbb{P}(E)$. 
**Example:** Suppose that an insurance company classifies people into three categories based on their annual risk of having an accident. Those in category 1 have a 1% risk, those in category 2 have a 5% risk, and those in category 3 have a 10% risk. If the frequencies of these three categories in a population are 70%, 20% and 10%, respectively, calculate the probability that a random sampled individual from this population will have an accident in a given year.

Solution: Let $C_1$, $C_2$, and $C_3$ denote the three risk classes and let $A$ be the event that a random sampled individual has an accident in a given year. Then, by the law of total probability,

$$P(A) = P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3) = 0.01 \times 0.7 + 0.05 \times 0.2 + 0.1 \times 0.1 = 0.027.$$
Example: Suppose that an insurance company classifies people into three categories based on their annual risk of having an accident. Those in category 1 have a 1% risk, those in category 2 have a 5% risk, and those in category 3 have a 10% risk. If the frequencies of these three categories in a population are 70%, 20% and 10%, respectively, calculate the probability that a random sampled individual from this population will have an accident in a given year.

Solution: Let $C_1$, $C_2$ and $C_3$ denote the three risk classes and let $A$ be the event that a random sampled individual has an accident in a given year. Then, by the law of total probability,

\[
P(A) = P(A|C_1)P(C_1) + P(A|C_2)P(C_2) + P(A|C_3)P(C_3)
\]

\[
= 0.01 \times 0.7 + 0.05 \times 0.2 + 0.1 \times 0.1
\]

\[
= 0.027.
\]
Earlier we saw that conditional probabilities are not symmetric, i.e., $P(A|B)$ and $P(B|A)$ are usually not equal. However, there is a simple relationship between these two probabilities which is known as Bayes' formula.

**Theorem**

(Bayes' formula) Suppose that $A$ and $B$ are events with $P(A) > 0$ and $P(B) > 0$. Then

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

**Proof:** Notice that we can write the joint probability of $A$ and $B$ in two different ways:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A).$$

Bayes' formula then follows from the equality of the two expressions on the right-hand side. $\square$
Example: Suppose that a test for HIV-1 infection is known to have a false positive rate of 2.3% and a false negative rate of 1.4%. If the prevalence of HIV-1 in a particular population is equal to 0.004, what is the probability that an individual with no known risk factors that tests positive for HIV-1 is, in fact, infected?

We can use Bayes’ formula to solve this problem, but we first introduce some notation. Let $H$ denote the proposition that the individual is infected and let $D$ denote the event that they test positive. Then,

$$
P(H|D) = P(H) \frac{P(D|H)}{P(D)},$$

and so we need to evaluate the three probabilities that appear on the right-hand side. First, since the individual has no known risk factors, we can think of them as a randomly sampled member of the population, which has prevalence 0.004. This suggests that we should make the following assignment:

$$P(H) = 0.004.$$
The quantity $P(D|H)$ is just the conditional probability that a person that is infected will test positive, which is determined by the false negative rate, i.e.,

$$P(D|H) = 1 - 0.014 = 0.986.$$ 

Lastly, a person can have a positive test result either because they are infected and correctly test positive or because they are not infected but receive a false positive. By the law of total probability, we know that

$$P(D) = P(D|H)P(H) + P(D|\overline{H})P(\overline{H}) = 0.986 \cdot 0.004 + 0.023 \cdot (1 - 0.004) = 0.02685.$$ 

Substituting these results back into Bayes’ formula shows that the probability that a random individual with a positive test result is actually infected is

$$P(H|D) = \frac{0.986}{0.02685} \approx 0.147.$$ 

Thus, despite the accuracy of the test, an individual with a positive test result but no other risk factors is not likely to be infected.
Example: In the US, approximately 8% of males and 0.5% of females have difficulty distinguishing between red and green hues. If an individual is sampled at random from this population and found to be color blind, what is the probability that they are male? (You may assume that the population contains an equal number of males and females.)

Solution:
Let $M$ and $F$ be the events that the sampled individual is male or female, respectively, and let $C$ be the event that a randomly sampled individual is red-green color blind. Then the probability that a randomly sampled individual who is color blind is also male is equal to

$$P(M|C) = P(M)P(C|M)P(C) = P(M)P(C|M)P(M) + P(C|F)P(F).$$

Approximately $0.08 * 0.5 + 0.005 * 0.5 = 0.0425 + 0.0025 = 0.045$.
Example: In the US, approximately 8% of males and 0.5% of females have difficulty distinguishing between red and green hues. If an individual is sampled at random from this population and found to be color blind, what is the probability that they are male? (You may assume that the population contains an equal number of males and females.)

Solution: Let $M$ and $F$ be the events that the sampled individual is male or female, respectively, and let $C$ be the event that a randomly sampled individual is red-green color blind. Then the probability that a randomly sampled individual who is color blind is also male is equal to

$$P(M|C) = \frac{P(M)P(C|M)}{P(C)}$$

$$= \frac{P(M)P(C|M)}{P(C|M)P(M) + P(C|F)P(F)}$$

$$= \frac{0.5 \times 0.08}{0.08 \times 0.5 + 0.005 \times 0.5}$$

$$\approx 0.9412.$$
Bayesian Inference

Bayes’ formula also provides us with a very flexible and powerful approach to statistical inference. Suppose that our goal is to use some new evidence to evaluate a set of competing hypotheses, $H_1, \cdots, H_n$. In Bayesian inference, we would perform the following three steps.

1. We first choose a prior distribution $(p_1, \cdots, p_n)$ on the set of hypotheses which quantifies how strongly we believe in each hypothesis before we examine the new data.

2. We then examine the new evidence $E$.

3. In light of this evidence, we revise our beliefs in the different hypotheses using Bayes’ formula:

$$p_i^* \equiv P(H_i | E) = P(H_i) \frac{P(E | H_i)}{P(E)} = \frac{P(E | H_i) p_i}{\sum_{j=1}^{n} P(E | H_j) p_j}.$$

The distribution $(p_1^*, \cdots, p_n^*)$ is called the posterior distribution on the hypotheses.
Example: Suppose that we wish to estimate the prevalence of an infection in a population containing \( n \) individuals. Unless \( n \) is small, we won’t be able to test every individual and so our estimate will have to rely on a sample of individuals from the population.

Let \( H_k \) be the hypothesis that \( k \) individuals in the population are infected and suppose that our prior distribution on \((H_0, \ldots, H_n)\) is uniform:

\[
p_k = \frac{1}{n+1}, \quad k = 0, 1, \ldots, n.
\]

Such a prior is said to be uninformative because it treats every hypothesis as equally likely, i.e., we have no prior information to favor one hypothesis over any of the others.

Now, suppose (for simplicity) that we sample a single member of the population and test them for the infection. Let \( E \) be the event that the individual is infected and notice that

\[
\mathbb{P}(E|H_k) = \frac{k}{n}
\]

since the probability of sampling an infected individual in a population in which \( k \) of \( n \) individuals are infected is just \( k/n \).
Furthermore, by the law of total probability, we know that

\[ P(E) = \sum_{k=0}^{n} P(E|H_k)P(H_k) = \sum_{k=0}^{n} \frac{k}{n(n+1)} = \frac{1}{n(n+1)} \frac{n(n+1)}{2} = \frac{1}{2}. \]

Using Bayes’ formula, the posterior probability of hypothesis \( H_i \) is

\[ p_i^* = p_i \frac{P(E|H_i)}{P(E)} = \frac{1}{n+1} \frac{i/n}{1/2} = \frac{2i}{n(n+1)}. \]

In particular, \( p_i^* > p_i \) whenever \( i > n/2 \). This is because we are more likely to sample an infected individual if the prevalence is greater than 1/2 than if it is less than 1/2 and so the evidence favors those hypotheses with prevalence greater than 1/2.
Independence

In general, there is no simple relationship between the probabilities $P(A)$, $P(B)$ and $P(AB)$. However, there is an important special case where the three probabilities are related by a simple identity.

**Definition**

- Events $A$ and $B$ are independent if $P(AB) = P(A)P(B)$.
- Events $A_1, A_2, \cdots, A_n$ are independent if for every $m \leq n$ and every collection $\{A_{i_1}, A_{i_2}, \cdots, A_{i_m}\}$ of distinct events the following identity is satisfied:

$$P \left( \bigcap_{j=1}^{m} A_{i_j} \right) = \prod_{j=1}^{m} P(A_{i_j})$$
Example: Suppose that a coin is tossed two times and assume that the probability of each of the four possible outcomes \( S = \{HH, HT, TH, TT\} \) is \(1/4\). Let \( A = \{HH, HT\} \) be the event that first toss results in heads and let \( B = \{HH, TH\} \) be the event that the second toss results in heads. Then \( AB = \{HH\} \) is the event that both tosses land on heads and so

\[
\mathbb{P}(A) = \frac{1}{2} \\
\mathbb{P}(B) = \frac{1}{2} \\
\mathbb{P}(AB) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B)
\]

which demonstrates that \( A \) and \( B \) are independent events.
In everyday speech, we would say that two events are independent if there is no causal or logical relationship between them, i.e., knowing that $B$ is true does not change the likelihood that $A$ is true. Indeed, if $A$ and $B$ are independent and $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$, then

$$
\begin{align*}
\mathbb{P}(A|B) &= \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(B)} = \mathbb{P}(A) \\
\mathbb{P}(B|A) &= \frac{\mathbb{P}(AB)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A)\mathbb{P}(B)}{\mathbb{P}(A)} = \mathbb{P}(B).
\end{align*}
$$

In other words, if $A$ and $B$ are independent, then conditioning on either event does not alter the probability of the other event. The reason that independence was defined using products rather than conditional probabilities is that this definition covers cases where one or more of the events has probability equal to 0. For example, the empty set is independent of every other event $A$ since

$$
\mathbb{P}(\emptyset \cap A) = \mathbb{P}(\emptyset) = 0 = \mathbb{P}(\emptyset)\mathbb{P}(A)
$$

no matter how $A$ is defined.
Exercise: Suppose that $A_1, \cdots, A_n$ are independent events. Show that

$$
\mathbb{P} \left( \bigcap_{i=1}^{m} A_i \bigg\vert \bigcap_{i=m+1}^{n} A_i \right) = \mathbb{P} \left( \bigcap_{i=1}^{m} A_i \right)
$$

for every $m = 1, \cdots, n - 1$. 
Exercise: Suppose that $A_1, \cdots, A_n$ are independent events. Show that

$$
P \left( \bigcap_{i=1}^{m} A_i \bigg| \bigcap_{i=m+1}^{n} A_i \right) = P \left( \bigcap_{i=1}^{m} A_i \right)
$$

for every $m = 1, \cdots, n - 1$.

Solution: Using the definitions of independence and conditional probabilities, we have

$$
P \left( \bigcap_{i=1}^{m} A_i \bigg| \bigcap_{i=m+1}^{n} A_i \right) = \frac{P \left( \bigcap_{i=1}^{n} A_i \right)}{P \left( \bigcap_{i=m+1}^{n} A_i \right)}
$$

$$
= \frac{\prod_{i=1}^{n} P(A_i)}{\prod_{i=m+1}^{n} P(A_i)}
$$

$$
= \prod_{i=1}^{m} P(A_i)
$$

$$
= P \left( \bigcap_{i=1}^{m} A_i \right) \Box
$$
Lemma

If the events $A$ and $B$ are independent, then so are the events $A$ and $\overline{B}$, $\overline{A}$ and $B$, and $\overline{A}$ and $\overline{B}$.

Proof: Since $A$ and $B$ are independent, we know that $P(AB) = P(A)P(B)$. Consequently,

$$P(A\overline{B}) = P(A) - P(AB)$$
$$= P(A) - P(A)P(B)$$
$$= P(A)(1 - P(B))$$
$$= P(A)P(\overline{B}),$$

which shows that $A$ and $\overline{B}$ are independent. This immediately implies that the other two pairs of events are also independent. 

Interpretation: To say that $A$ and $B$ are independent means that we learn nothing about $B$ upon being told that $A$ is true. However, in that case, we also learn nothing about $B$ if we are told that $A$ is false, i.e., that $\overline{A}$ is true. Thus $\overline{A}$ and $B$ must also be independent.
**Example:** Suppose that a fair die is rolled twice and that the probability of each of the 36 possible outcomes \{ (1, 1), (1, 2), \cdots, (6, 6) \} is equal to 1/36. For \( 1 \leq i, j \leq 6 \), let \( A_i \) be the event that the first roll lands on the number \( i \) and let \( B_j \) be the event that the second roll lands on the number \( j \). Then \( A_i B_j \) is the event that we first roll an \( i \) and then a \( j \), and so

\[
\begin{align*}
\mathbb{P}(A_i) &= \frac{1}{6} \\
\mathbb{P}(B_j) &= \frac{1}{6} \\
\mathbb{P}(A_i B_j) &= \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A_i)\mathbb{P}(B_j)
\end{align*}
\]

which demonstrates that \( A_i \) and \( B_j \) are independent events for every pair \( i \) and \( j \). Informally, we say that the first and the second rolls of the die are independent.
Continuing with the example from the preceding slide, let

\[ C = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} \]

be the event that the sum of the two numbers rolled is equal to 7. Since \( C \) contains 6 possible outcomes, each having probability \( \frac{1}{36} \), we see that \( \mathbb{P}(C) = \frac{1}{6} \). Also, the events \( A_1 C = \{(1, 6)\} \) and \( B_1 C = \{(6, 1)\} \) each contain one outcome and so

\[
\begin{align*}
\mathbb{P}(A_1 C) &= \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(A_1)\mathbb{P}(C) \\
\mathbb{P}(B_1 C) &= \frac{1}{36} = \frac{1}{6} \times \frac{1}{6} = \mathbb{P}(B_1)\mathbb{P}(C),
\end{align*}
\]

which shows that the events \( A_1 \) and \( C \) are independent, as are \( B_1 \) and \( C \). In other words, being told that the sum of the two numbers rolled is equal to 7 provides us with no new information concerning the first number rolled or concerning the second number rolled.
Now, consider all three events $A_1, B_1$ and $C$ together. Clearly $A_1 B_1 C = \emptyset$, since if the first and the second numbers rolled are both equal to 1, then their sum cannot be equal to 7. Consequently,

$$0 = P(A_1 B_1 C) \neq P(A_1)P(B_1)P(C) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216}.$$

This shows that although $A_1, B_1$ and $C$ are pairwise independent, i.e., each pair of events is independent, the three events considered together are not independent. In other words, simply knowing that one of these three events occurred tells us nothing about any one of the remaining two events, but knowing that two of these events has occurred does provide us with information about the third event.

**Moral:** Independence implies pairwise independence, but pairwise independence does not imply independence.
In certain instances, two events that are dependent may become independent when we are provided with some additional information. This possibility is the subject of the following definition.

**Definition**

Suppose that $B$ is an event with $\mathbb{P}(B) > 0$. The events $A_1, \cdots, A_n$ are said to be **conditionally independent given $B$** if for every $m \leq n$ and every collection $\{A_{i_1}, \cdots, A_{i_m}\}$ of distinct events the following identity is satisfied:

$$\mathbb{P}
\left(
\bigcap_{j=1}^{m}
A_{i_j} \bigg| B
\right) = \prod_{j=1}^{m} \mathbb{P}(A_{i_j} | B).$$
Example: Suppose that a jar contains two biased coins, one (coin 1) having probability $1/4$ of landing on heads and the other (coin 2) having probability $3/4$ of landing on heads. A coin is chosen at random from this jar and flipped twice. There are eight possible outcomes for this experiment, which we can represent by the following sample space:

$$S = \{(1HH), (1HT), (1TH), (1TT), (2HH), (2HT), (2TH), (2TT)\}.$$  

For example, the outcome $(1HH)$ means that coin 1 was sampled and that both tosses landed on heads.

Suppose that the probabilities of these eight outcomes are given by the numbers in the following table:

<table>
<thead>
<tr>
<th></th>
<th>$1HH$</th>
<th>$1/32$</th>
<th>$2HH$</th>
<th>$9/32$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$1HT$</td>
<td>$3/32$</td>
<td>$2HT$</td>
<td>$3/32$</td>
</tr>
<tr>
<td></td>
<td>$1TH$</td>
<td>$3/32$</td>
<td>$2TH$</td>
<td>$3/32$</td>
</tr>
<tr>
<td></td>
<td>$1TT$</td>
<td>$9/32$</td>
<td>$2TT$</td>
<td>$1/32$</td>
</tr>
</tbody>
</table>

Indeed, since these numbers are non-negative and sum to one, they define a proper probability distribution on $S$. 
Let $A_1$ be the event that the first toss lands on heads, let $A_2$ be the event that the second toss lands on heads, and let $B$ be the event that coin 1 is chosen. I claim that $A_1$ and $A_2$ are conditionally independent given $B$. To see that this is true, observe that

\[
P(B) = P(1HH, 1HT, 1TH, 1TT) = \frac{1}{2}
\]

\[
P(A_1|B) = \frac{P(1HH, 1HT)}{P(B)} = \frac{4/32}{1/2} = \frac{1}{4}
\]

\[
P(A_2|B) = \frac{P(1HH, 1TH)}{P(B)} = \frac{4/32}{1/2} = \frac{1}{4}
\]

\[
P(A_1A_2|B) = \frac{P(1HH)}{P(B)} = \frac{1/32}{1/2} = \frac{1}{16} = P(A_1|B)P(A_2|B).
\]

Similarly, $A_1$ and $A_2$ are conditionally independent given $\overline{B}$ since

\[
P(A_1A_2|\overline{B}) = \frac{9}{16} = \frac{3}{4} \cdot \frac{3}{4} = P(A_1|\overline{B})P(A_2|\overline{B}).
\]
On the other hand, the events $A_1$ and $A_2$ are not independent. Indeed,

\[
\begin{align*}
P(A_1) &= P(A_1|B)P(B) + P(A_1|\overline{B})P(\overline{B}) \\
&= \frac{1}{4} \times \frac{1}{2} + \frac{3}{4} \times \frac{1}{2} = \frac{1}{2}
\end{align*}
\]

\[
\begin{align*}
P(A_2) &= P(A_2|B)P(B) + P(A_2|\overline{B})P(\overline{B}) \\
&= \frac{1}{4} \times \frac{1}{2} + \frac{3}{4} \times \frac{1}{2} = \frac{1}{2}
\end{align*}
\]

\[
\begin{align*}
P(A_1A_2) &= P(A_1A_2|B)P(B) + P(A_1A_2|\overline{B})P(\overline{B}) \\
&= \frac{1}{16} \times \frac{1}{2} + \frac{9}{16} \times \frac{1}{2} = \frac{5}{16}
\end{align*}
\]

\[
\ne \neq \frac{1}{4} = P(A_1)P(A_2).
\]

**Explanation:** To understand why two events are not independent, we should consider what we learn about one event when we are told that the other event is true. In this case, being told that $A_1$ has occurred makes it more likely that we have sampled coin 2 because this coin has a higher probability of landing on heads. Having learned this information, we are then in a better position to predict whether $A_2$ will occur, since the probability of this event depends on which coin we have sampled.
Suppose that there are two routes up a mountain, one easy and the other difficult, and also that there are two classes of climbers that differ in experience, amateurs and professionals. For convenience, we introduce the following notation:

- $D$ is the event that a climber attempts the difficult route;
- $A$ is the event that the climber is an amateur;
- $S$ is the event that the climber reaches the summit.

We will assume that professional climbers are more likely to reach the summit no matter which route they attempt:

$$P(S|D, A) = 0.3$$  $$P(S|D, \overline{A}) = 0.4$$
$$P(S|\overline{D}, A) = 0.7$$  $$P(S|\overline{D}, \overline{A}) = 0.8$$

We will also assume that professional climbers are more likely to attempt the difficult route:

$$P(D|A) = 0.35$$  $$P(D|\overline{A}) = 0.75.$$
Then the overall probability that an amateur climber will reach the summit, irrespective of which route they attempt, is

\[ P(S|A) = P(S|D, A)P(D|A) + P(S|\overline{D}, A)P(\overline{D}|A) \]
\[ = 0.3 \times 0.35 + 0.7 \times 0.65 = 0.56. \]

Similarly, the overall probability that a professional climber will reach the summit is

\[ P(S|A) = P(S|D, \overline{A})P(D|\overline{A}) + P(S|\overline{D}, \overline{A})P(\overline{D}|\overline{A}) \]
\[ = 0.4 \times 0.75 + 0.8 \times 0.25 = 0.5. \]

Thus amateur climbers are more likely to reach the summit than professional climbers despite the fact that they are less likely to be successful along either route. The reason for this, of course, is that amateurs are more likely than professionals to attempt the easier route. However, if we were not aware of the different routes and only considered the aggregated data, then we might wrongly conclude that amateurs were better climbers than professionals. This is the essence of Simpson’s paradox.
Here is an example using real data (Julious & Mullee, BMJ 1994). A comparative study of two surgical procedures for kidney stone removal found the following success rates:

<table>
<thead>
<tr>
<th></th>
<th>open surgery</th>
<th>percutaneous nephrolithotomy</th>
</tr>
</thead>
<tbody>
<tr>
<td>small stones</td>
<td>93% (81/87)</td>
<td>87% (234/270)</td>
</tr>
<tr>
<td>large stones</td>
<td>73% (192/263)</td>
<td>69% (55/80)</td>
</tr>
<tr>
<td>aggregated</td>
<td>78% (273/350)</td>
<td>83% (289/350)</td>
</tr>
</tbody>
</table>

Thus open surgery appears to be more successful when the data from either the small stone cases or the large stone cases is analyzed separately, but percutaneous nephrolithotomy appears more successful when all of the data is analyzed together. There are two reasons for this:

- The size of the stone has a large impact on the success of the treatment, whichever method is used. In particular, outcomes are poorer for large stones than for small stones.
- The proportions of cases with large or small stone sizes differ between the two treatments. Specifically, percutaneous nephrolithotomy was used much more frequently to treat small kidney stones.
In this example, kidney stone size is a **confounding variable**, i.e., a variable that is correlated with both the treatment and the effect. Failure to account for confounding variables can lead to spurious or incorrect results. For example, had kidney stone size been neglected in this study, then the authors might have wrongly concluded that percutaneous nephrolithotomy is the more effective treatment. The difficulty is that we can never be certain that we have accounted for all of the variables that could be confounding our analysis. We can attempt to guard against this danger in two ways:

- **Randomization:** Test subjects should be randomly assigned to treatment groups. This will help to reduce the correlations between the treatment and any unmeasured covariates.

- **Expert knowledge:** It is important to be aware of possible causal relationships that might exist between the variables of interest and other confounding variables. This depends on having a detailed understanding of the system being studied.
Suppose that two opponents, $A$ and $B$, engage in a series of games with the following characteristics:

- Each game ends in a victory for one of the two players (i.e., there are no draws) and player $A$ has probability of winning any particular game, while player $B$ has probability $q = 1 - p$ of winning that game.
- The outcomes of the games are independent of one another, e.g., the fact that player $A$ wins the first and second games makes it neither more nor less likely that they will win the third game.
- Each time that a player loses a game, they give a dollar to their opponent.
- Player $A$ begins the contest with $i$ dollars, while player $B$ begins with $n - i$ dollars.
- The contest ends as soon as one of the two players has no money left; that player is said to be ruined and the other player is deemed to have won the contest.

**Problem:** What is the probability that $A$ will win the contest?
The gambler’s ruin problem has a long history, having been posed at least as early as 1656 by Blaise Pascal and solved soon after by Christiaan Huygens. Let $W_i$ denote the event that $A$ wins the contest when they start with $i$ dollars and let $a_i = \mathbb{P}(W_i)$ be the probability of this outcome.

Since it isn’t obvious how to calculate $a_i$, we need to find some additional information that we can condition on to simplify the problem. In this case, the contest consists of a series of games with independent outcomes and so one candidate for this information is the outcome of the first game. If we let $G$ denote the event that $A$ is the winner of the first game, then by the law of total probability we know that

$$a_i \equiv \mathbb{P}(W_i) = \mathbb{P}(W_i | G)\mathbb{P}(G) + \mathbb{P}(W_i | \overline{G})\mathbb{P}(\overline{G}) = p \cdot \mathbb{P}(W_i | G) + q \cdot \mathbb{P}(W_i | \overline{G}),$$

since $p = \mathbb{P}(G)$ and $q = \mathbb{P}(\overline{G})$. Of course, this will be useful only if we can calculate the conditional probabilities $\mathbb{P}(W_i | G)$ and $\mathbb{P}(W_i | \overline{G})$. 

I claim that

\[ \mathbb{P}(W_i | G) = \mathbb{P}(W_{i+1}) \equiv a_{i+1} \]
\[ \mathbb{P}(W_i | \overline{G}) = \mathbb{P}(W_{i-1}) \equiv a_{i-1}. \]

The first identity is a consequence of the following logic:

- If \( A \) starts with \( i \) dollars and wins the first game, i.e., \( G \) is true, then after the first game, \( A \) will possess \( i + 1 \) dollars while \( B \) will possess \( n - i - 1 \) dollars.
- The outcomes of the second and all subsequent games are independent of both the outcome of the first game and the amount of money that each player had at the start of the contest.
- From the second game forwards, it is as if the players are starting an entirely new contest in which player \( A \) begins with \( i + 1 \) dollars and player \( B \) begins with \( n - i - 1 \) dollars.
- By definition, the probability that \( A \) wins this new contest is \( a_{i+1} \) and so \( \mathbb{P}(W_i | G) = a_{i+1} \).

Since \( A \) will only have \( i - 1 \) dollars if \( \overline{G} \) is true, similar reasoning shows that \( \mathbb{P}(W_i | \overline{G}) = a_{i-1} \).
Combining the results from the previous two slides and noting that \( A \) cannot win if they begin with no money and they are certain to win if the begin with all of the money, we arrive at the following system of equations for the \( a_i \)'s:

\[
\begin{align*}
a_0 &= 0 \\
a_n &= 1 \\
a_i &= p \cdot a_{i+1} + q \cdot a_{i-1}, \quad 1 \leq i \leq n - 1.
\end{align*}
\]

This is an example of a \textbf{two term recursion}, since each value \( a_{i+1} \) depends on the two preceding values \( a_i \) and \( a_{i-1} \). To solve this, we begin by subtracting \( a_i \) from both sides of the last equation:

\[
\begin{align*}
0 &= p \cdot a_{i+1} + q \cdot a_{i-1} - a_i \\
&= p \cdot a_{i+1} + q \cdot a_{i-1} - (p + q) \cdot a_i \\
&= p \cdot (a_{i+1} - a_i) - q \cdot (a_i - a_{i-1}) \\
&= p \cdot \Delta_i - q \cdot \Delta_{i-1},
\end{align*}
\]

where we have introduce a new set of variables \( \Delta_i \equiv a_{i+1} - a_i, \ 0 \leq i \leq n - 1 \).
The last equation on the preceding slide is an example of a **one term recursion**, which can be rewritten as:

\[ \Delta_i = \left( \frac{q}{p} \right) \Delta_{i-1} = \alpha \Delta_{i-1} \]

where we define \( \alpha \equiv q/p \). This recursion can be solved iteratively:

\[
\begin{align*}
    a_2 - a_1 &= \Delta_1 = \alpha \Delta_0 \\
    a_3 - a_2 &= \Delta_2 = \alpha \Delta_1 = \alpha^2 \Delta_0 \\
    a_4 - a_3 &= \Delta_3 = \alpha \Delta_2 = \alpha^3 \Delta_0 \\
    & \quad \ldots \\
    a_{i+1} - a_i &= \Delta_i = \alpha^i \Delta_0
\end{align*}
\]

for \( i = 0, \ldots, n - 1 \). This is progress since we now know how the \( \Delta_i \)'s are related to one another, but we still need to use this result to solve for the values of \( a_1, \ldots, a_{n-1} \).
The \( a_i \)'s can be expressed in terms of the \( \Delta_i \)'s in the following manner. First, because \( a_0 = 0 \), we know that

\[
\Delta_0 = a_1 - a_0 = a_1.
\]

We can then solve for \( a_i \) with the help of a telescoping sum:

\[
a_i = (a_i - a_{i-1}) + (a_{i-1} - a_{i-2}) + \cdots + (a_2 - a_1) + a_1
\]

\[
= \Delta_{i-1} + \Delta_{i-2} + \cdots + \Delta_1 + \Delta_0
\]

\[
= \alpha^{i-1} \Delta_0 + \alpha^{i-2} \Delta_0 + \cdots + \alpha \Delta_0 + \Delta_0
\]

\[
= \left( \sum_{k=0}^{i-1} \alpha^k \right) \Delta_0
\]

\[
= \begin{cases} 
\left( \frac{1 - \alpha^i}{1 - \alpha} \right) a_1 & \text{if } \alpha \neq 1 \\
\alpha a_1 & \text{if } \alpha = 1.
\end{cases}
\]
Remark: The last result was obtained with the help of the following formula for the sum of a geometric series with finitely many terms:

\[
\sum_{i=0}^{n} r^i = \begin{cases} 
\frac{1-r^{n+1}}{1-r} & \text{if } r \neq 1 \\
n + 1 & \text{if } r = 1.
\end{cases}
\]

Indeed, for any value of \( r \), we have

\[
(1 - r) \sum_{i=0}^{n} r^i = (1 - r) \left(1 + r + r^2 + \cdots + r^n\right)
\]

\[
= 1 + r + r^2 + \cdots + r^{n-1} + r^{n} - r - r^2 - r^3 - \cdots - r^n - r^{n+1}
\]

\[
= 1 - r^{n+1},
\]

and so the topmost identity follows upon dividing by \( 1 - r \), which we can do so long as \( r \neq 1 \).
We still need to solve for \( a_1 \), which we can do with the help of the remaining boundary condition \( a_n = 1 \). Taking \( i = n \) in the expressions derived previously gives

\[
1 = a_n = \begin{cases} 
\left( \frac{1 - \alpha^n}{1 - \alpha} \right) a_1 & \text{if } \alpha \neq 1 \\
n a_1 & \text{if } \alpha = 1,
\end{cases}
\]

which shows that

\[
a_1 = \begin{cases} 
\left( \frac{1 - \alpha}{1 - \alpha^n} \right) & \text{if } \alpha \neq 1 \\
1/n & \text{if } \alpha = 1.
\end{cases}
\]
Finally, upon substituting these results back into our previous expressions for $a_i$, we arrive at the solution to the gambler’s ruin problem:

$$a_i = \begin{cases} 
\left( \frac{1-\alpha^i}{1-\alpha^n} \right) & \text{if } \alpha \neq 1 \\
i/n & \text{if } \alpha = 1.
\end{cases}$$

for $i = 0, \cdots, n$. In particular, if the two players are evenly matched, i.e., $p = q = 1/2$, then $\alpha = 1$ and so a player’s chances of winning all of the money are equal to the proportion of the money that the player starts with.

Although the earliest formulation of the gambler’s ruin problem was motivated by interest in games of chance, its solution is pertinent to problems in other disciplines. In the next several slides I give an example of an ecological model which is concerned with the effects of competition on the species composition of an ecological community.
Suppose that two species, $A$ and $B$, are competing for a common pool of resources in a habitat that will support at most $n$ adult individuals. We will make the following assumptions:

- Individuals of both species die at the same rate.
- Reproduction only occurs when the death of an individual has opened up space in the habitat for a new individual.
- Individuals of species $A$ reproduce at rate $s_A$, while those of species $B$ reproduce at rate $s_B$.
- If the population contains $i$ individuals of species $A$ and $n - i$ individuals of species $B$, then when a death occurs, the dying individual is replaced by a new individual of species $A$ with probability

$$\pi_i = \frac{is_A}{is_A + (n - i)s_B}$$

Otherwise, the dying individual is replaced by a new individual of species $B$.

In this case, we would like to calculate the probability that species $A$ will eventually replace species $B$; this is also the probability of extinction of species $B$. 
The process described on the previous slide is sometimes known as the **Moran model**, after P.A.P. Moran, an Australian mathematical biologist. It is related to the gambler’s ruin in the following way.

Suppose that the population contains \( i \) individuals of species \( A \). After the next birth-death event, the population will either have \( i - 1 \), \( i \) or \( i + 1 \) individuals of this species:

- If a member of species \( B \) dies and is replaced by a member of species \( A \), then the population size of \( A \) will increase to \( i + 1 \).
- If a member of species \( A \) dies and is replaced by a member of species \( B \), then the population size of \( A \) will decrease to \( i - 1 \).
- If the dying individual and their replacement belong to the same species, either \( A \) or \( B \), then the number of individuals of each species will remain unchanged.

Let us calculate the probabilities of these three events. We will use the notation \( x_i = i/n \) to denote the proportion of the population made up of members of species \( A \).
We begin with the case in which the population size of species $A$ increases from $i$ to $i+1$. For this to occur, a member of species $B$ must die and be replaced by a member of species $A$. I claim that the probability of this event is $\pi_i(1 - x_i)$.

- First, the probability that the dying individual belongs to species $B$ is $1 - x_i$. Indeed, since both species die at equal rates, each of the $n$ individuals in the population is equally likely to die and thus the probability that the dying individual belongs to species $B$ is equal to the proportion of individuals in the population that belong to this species, which is just $1 - p_i$.

- The probability that the newborn individual belongs to species $A$ was previously defined to be $\pi_i = is_A/(is_A + (n - i)s_B)$.

- Since the species of the dying individual and the newborn individual are independent, the probability of the birth and the death being as specified is just the product of these two probabilities.
By similar arguments,

- The probability that the number of members of species $A$ decreases by one is $(1 - \pi_i)x_i$.
- Since the probabilities of an increase, a decrease, or no change in size must sum to 1, we can conclude that the probability that the numbers of neither species change is equal to $1 - \pi_i(1 - x_i) - (1 - \pi_i)x_i$.

Notice that events of the third type do not change the composition of the population and thus do not affect the probability of extinction of species $B$. These events can be ignored in our current analysis. On the other hand, events that lead to changes in the population sizes of the two species will have a direct influence on which species eventual goes extinct.
With these considerations in mind, let us calculate the conditional probability that the number of individuals of species $A$ increases by one given that there is a change in the population size of this species. Let $I$ be the event that the population of species $A$ increases by one and let $D$ be the event that it decreases by one. Then $I \cup D$ is the event that this population changes and so the probability that concerns us is

$$
P(I|I \cup D) = \frac{P(I)}{P(I \cup D)} = \frac{P(I)}{P(I) + P(D)} = \frac{\pi_i (1 - x_i)}{\pi_i (1 - x_i) + (1 - \pi_i) x_i} = \frac{\frac{i s_A}{i s_A + (n-i) s_B} \frac{n-i}{n}}{\frac{i s_A}{i s_A + (n-i) s_B} \frac{n-i}{n} + \frac{(n-i) s_B}{i s_A + (n-i) s_B} \frac{i}{n}} = \frac{i(n-i)s_A}{i(n-i)(s_A + s_B)} = \frac{s_A}{s_A + s_B}.
$$
Thus the conditional probability of an increase in population \( A \) by one given any change at all does not depend on the current size of the population and is equal to \( p = s_A/(s_A + s_B) \). Similarly, the conditional probability of a decrease in population \( S \) by one given any change at all also does not depend on the current population size and is equal to \( q = s_B/(s_A + s_B) \).

It follows that if we restrict attention to events that change the composition of the population, then we are exactly faced with the gambler’s ruin problem and so we can use our solution to that problem to address this one as well. Letting

\[
\alpha = \frac{q}{p} = \frac{s_B}{s_A},
\]

the probability that species \( A \) will eventually replace species \( B \) is

\[
a_i = \begin{cases} 
\left(\frac{1-\alpha^i}{1-\alpha^n}\right) & \text{if } s_A \neq s_B \\
i/n & \text{if } s_A = s_B.
\end{cases}
\]
These calculations lead to the following conclusions:

- If both species are equally matched, i.e., if $s_A = s_B$, then the probability that one species eventually displaces the other is equal to the initial frequency of that species.

- If one species is competitively superior to the other species, then that species will be disproportionately likely to drive the other to extinction. For example, if $A$ reproduces more rapidly than $B$, then $s_A > s_B$ and $a_i > i/n$.

- However, even if one species enjoys a competitive advantage, it may still be driven to extinction by the other species just by chance. This is more likely to happen in smaller populations.
**Problem:** For another example of how conditioning on the past can be used to calculate the probabilities of future events, consider the following simple model of the weather. Suppose that days are characterized as either wet or dry and that the probability that the weather tomorrow is the same as the weather today is $p$ (some constant in $(0, 1)$). Show that if the weather is dry today, then the probability $P_n$ that the weather is dry $n$ days from now is given by the following formula:

$$P_n = \frac{1}{2} + \frac{1}{2}(2p - 1)^n$$

for $n \geq 0$. 
**Solution:** Let $E_n$ be the event that the weather is dry $n$ days from now and let $P_n = \mathbb{P}(E_n)$. By conditioning on the previous day’s weather, we have

\[
P_n = \mathbb{P}(E_n) = \mathbb{P}(E_n|E_{n-1})\mathbb{P}(E_{n-1}) + \mathbb{P}(E_n|E_{n-1}^c)\mathbb{P}(E_{n-1}^c)
\]

\[
= pP_{n-1} + (1 - p)(1 - P_{n-1})
\]

\[
= (2p - 1)P_{n-1} + (1 - p),
\]

with $P_0 = 1$ by assumption. Direct substitution then shows that

\[
P_n = \frac{1}{2} + \frac{1}{2}(2p - 1)^n
\]

is the unique solution to this recursion satisfying $P_0 = 1$. 

Jay Taylor (ASU) Fall 2013 63 / 133
Example: Suppose that an individual is sampled at random from a population and let $A$ be the age of the individual in years and let $G \in \{f, m\}$ denote the gender of the individual. The **joint distribution** of $A$ and $G$, consisting of the probabilities $p_{a,g} = \mathbb{P}(A = a, G = g)$, is shown in the following table.

<table>
<thead>
<tr>
<th>$p_{a,g}$</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>0.018</td>
<td>0.061</td>
<td>0.101</td>
<td>0.053</td>
<td>0.121</td>
<td>0.115</td>
</tr>
<tr>
<td>$m$</td>
<td>0.106</td>
<td>0.121</td>
<td>0.099</td>
<td>0.107</td>
<td>0.083</td>
<td>0.015</td>
</tr>
</tbody>
</table>

Use this table to calculate the following quantities:

1. the probability that the individual is female, $\mathbb{P}(G = f)$;
2. the probability that their age is 20, $\mathbb{P}(A = 20)$;
3. the expected age of the individual, $\mathbb{E}[A]$;
4. the expected age of the individual, given that they are female;
5. the expected age of the individual, given that they are male;
Solutions:

1.) Since \( \{G = f\} = \{(17, f), (18, f), (19, f), (20, f), (21, f), (22, f)\} \), it follows that
\[
P(G = f) = 0.120 + 0.061 + 0.101 + 0.018 + 0.053 + 0.115 = 0.469.
\]

2.) Since \( \{A = 20\} = \{(20, f), (20, m)\} \), it follows that
\[
P(A = 20) = 0.053 + 0.107 = 0.160.
\]

3.) Using the definition of expectation, the expected age of a randomly selected individual is
\[
\mathbb{E}[A] = \sum_{a=17}^{23} P(A = a) \cdot a = \sum_{a=17}^{23} (p_{a,f} + p_{a,m}) \cdot a
\]
\[
= (0.018 + 0.106) \cdot 17 + (0.061 + 0.121) \cdot 18 + (0.101 + 0.099) \cdot 19 +
\]
\[
(0.053 + 0.107) \cdot 20 + (0.121 + 0.083) \cdot 21 + (0.115 + 0.015) \cdot 22
\]
\[
= 19.528.
\]
4.) To find the expected age of an individual given that they are female, we need to calculate the **conditional age distribution** of the female members of the population. This is the distribution made up of the conditional probabilities of the ages given that the randomly chosen individual is female:

\[ p_{a|f} = \mathbb{P}(A = a | G = f) = \frac{\mathbb{P}(A = a, G = f)}{\mathbb{P}(G = f)} = \frac{p_{a,f}}{0.468} \]

and the actual values are shown in the following table:

| \( p_{a|f} \) | 17   | 18   | 19   | 20   | 21   | 22   |
|----------------|------|------|------|------|------|------|
| \( f \)       | 0.0384 | 0.1301 | 0.2154 | 0.113 | 0.258 | 0.2452 |

Then, the expected age of an individual given that they are female is equal to the expectation of the conditional age distribution:

\[
\mathbb{E}[A|G = f] = \sum_{a=17}^{22} p_{a|f} \cdot a \approx 20.1578.
\]
5.) Similarly, the conditional age distribution of the male members of the population is given by the values

\[ p_{a|m} = \frac{p_{a,m}}{P(G = m)} = \frac{p_{a,m}}{0.532}, \]

which are shown in the following table:

| \( p_{a|m} \) | 17   | 18   | 19   | 20   | 21   | 22   |
|----------------|------|------|------|------|------|------|
| \( m \)       | 0.1996 | 0.2279 | 0.1864 | 0.2015 | 0.1563 | 0.0282 |

Using these values, we find that the expected age of a male member of the population is

\[ \mathbb{E}[A|G = m] = \sum_{a=17}^{23} P(A = a|G = m) \cdot a \approx 18.9718. \]
Having calculated three expectations of age,

\[
\begin{align*}
\mathbb{E}[A|G = f] &= 20.1578 \\
\mathbb{E}[A|G = m] &= 18.9718 \\
\mathbb{E}[A] &= 19.528
\end{align*}
\]

it is natural to ask how they are related. The first thing that we notice is that the average age of the population is intermediate between the average age of the females and the average age of the males. In fact, more than this is true. The average age of the population is equal to the weighted average of the average age of the males and the average age of the females, where the weights are equal to the frequencies of the two sexes in the population:

\[
\mathbb{E}[A] = \mathbb{P}(G = f) \cdot \mathbb{E}[A|G = f] + \mathbb{P}(G = m) \cdot \mathbb{E}[A|G = m].
\]

In other words, to calculate the average age of a random member of the population, we can first calculate the average ages of male and female members and then average these values using the frequencies of the two sexes. This is a special case of a general result that we will derive in the following slides.
Suppose that $X$ and $Y$ are random variables with values in the sets $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$, respectively, and let the joint distribution of $X$ and $Y$ be given by the values

$$p_{i,j} = P(X = x_i, Y = y_j), \quad 1 \leq i \leq n, 1 \leq j \leq m.$$ 

In other words, the joint distribution of two random variables tells us the probabilities of events that depend on the values of both random variables simultaneously. Recall that if we are given the joint distribution, then we can also calculate the marginal distribution of each random variable, which tells us the probabilities of events that depend on just one random variable:

$$p_i^X \equiv P(X = x_i) = \sum_{j=1}^{m} p_{i,j}$$

$$p_j^Y \equiv P(Y = y_j) = \sum_{i=1}^{n} p_{i,j}.$$
When solving problems involving two random variables, it may be that we know the value of one of the two variables, say $Y = y_j$, and we want to use this information to update our beliefs concerning the value of the other variable $X$. This consideration motivates the following definition.

**Definition**

Let $X$ and $Y$ be random variables with values in the sets $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$, respectively, and joint distribution $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$. Then the **conditional distribution** of $X$ given $Y = y_j$ is the distribution on the set $\{x_1, \cdots, x_n\}$ defined by

$$p_{i|j} \equiv \mathbb{P}(X = x_i | Y = y_j) = \frac{\mathbb{P}(X = x_i, Y = y_j)}{\mathbb{P}(Y = y_j)} = \frac{p_{i,j}}{p_j^Y}.$$

**Remark:** The conditional distribution of $X$ given $Y = y_j$ is a proper probability distribution satisfying all of the usual properties of probability distributions. For instance,

$$\sum_{i=1}^n p_{i|j} = \sum_{i=1}^n \frac{p_{i,j}}{p_j^Y} = \frac{1}{p_j^Y} \sum_{i=1}^n p_{i,j} = \frac{1}{p_j^Y} \cdot p_j^Y = 1.$$
Example: Suppose that a fair die is rolled twice and assume that the probability of each of the 36 possible outcomes \{ (1,1), (1,2), \cdots , (6,6) \} is equal to 1/36. Let \( X \) be the first number rolled, let \( Y \) be the second number rolled, and let \( S = X + Y \) be the sum of these two numbers. The conditional distribution of \( X \) given \( S = s \) is summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1/2</td>
<td>1/3</td>
<td>1/4</td>
<td>1/5</td>
<td>1/6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/2</td>
<td>1/3</td>
<td>1/4</td>
<td>1/5</td>
<td>1/6</td>
<td>1/5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/3</td>
<td>1/4</td>
<td>1/5</td>
<td>1/6</td>
<td>1/5</td>
<td>1/4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/4</td>
<td>1/5</td>
<td>1/6</td>
<td>1/5</td>
<td>1/4</td>
<td>1/3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/5</td>
<td>1/6</td>
<td>1/5</td>
<td>1/4</td>
<td>1/3</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/6</td>
<td>1/5</td>
<td>1/4</td>
<td>1/3</td>
<td>1/2</td>
<td>1</td>
</tr>
</tbody>
</table>

For example,

\[
P(X = 3 | S = 9) = \frac{P(X = 3, S = 9)}{P(S = 9)} = \frac{P(X = 3, Y = 6)}{P(S = 9)} = \frac{1/36}{4/36} = \frac{1}{4}.
\]
If $X$ takes on numeric values, then we can calculate both the expected value of $X$ and the expected value of $X$ given $Y = y_j$. The latter quantity is defined below.

**Definition**

Let $X$ and $Y$ be random variables with values in the sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$, respectively, and joint distribution $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$. Then the conditional expectation of $X$ given $Y = y_j$ is the quantity defined by

$$
\mathbb{E}[X|Y = y_j] = \sum_{i=1}^{n} p(X = x_i|Y = y_j) \cdot x_i = \frac{1}{p_j} \sum_{i=1}^{n} p_{i,j} \cdot x_i
$$

Conditional expectations have the same properties as ordinary expectations. For example, if we have three random variables $X_1$, $X_2$ and $Y$ and constants $c_1$ and $c_2$, then

$$
\mathbb{E}[c_1 X_1 + c_2 X_2|Y = y_j] = c_1 \mathbb{E}[X_1|Y = y_j] + c_2 \mathbb{E}[X_2|Y = y_j].
$$
Example, cont’d: The conditional expectation of $X$ given $S = s$ is:

$$
\begin{align*}
\mathbb{E}[X|S = 2] &= 1 \cdot 1 = 1 \\
\mathbb{E}[X|S = 3] &= \frac{1}{2} \cdot (1 + 2) = 1.5 \\
\mathbb{E}[X|S = 4] &= \frac{1}{3} \cdot (1 + 2 + 3) = 2 \\
\mathbb{E}[X|S = 5] &= \frac{1}{4} \cdot (1 + 2 + 3 + 4) = 2.5 \\
\mathbb{E}[X|S = 6] &= \frac{1}{5} \cdot (1 + 2 + 3 + 4 + 5) = 3 \\
\mathbb{E}[X|S = 7] &= \frac{1}{6} \cdot (1 + 2 + 3 + 4 + 5 + 6) = 3.5 \\
\mathbb{E}[X|S = 8] &= \frac{1}{5} \cdot (2 + 3 + 4 + 5 + 6) = 4 \\
\mathbb{E}[X|S = 9] &= \frac{1}{4} \cdot (3 + 4 + 5 + 6) = 4.5 \\
\mathbb{E}[X|S = 10] &= \frac{1}{3} \cdot (4 + 5 + 6) = 5 \\
\mathbb{E}[X|S = 11] &= \frac{1}{2} \cdot (5 + 6) = 5.5 \\
\mathbb{E}[X|S = 12] &= 1 \cdot 6 = 6.
\end{align*}
$$
We can also define the conditional expectation of $X$ given $Y$ when the value of $Y$ is unknown!

**Definition**

Let $X$ and $Y$ be random variables with values in the sets $\{x_1, \ldots, x_n\}$ and $\{y_1, \ldots, y_m\}$, respectively, and joint distribution $p_{ij} = \mathbb{P}(X = x_i, Y = y_j)$. Then the conditional expectation of $X$ given $Y$ is the random variable denoted $\mathbb{E}[X|Y]$ which is defined as follows:

$$\mathbb{E}[X|Y] \equiv \mathbb{E}[X|Y = y_j] \text{ whenever } Y = y_j.$$ 

Notice that $\mathbb{E}[X|Y]$ is defined as a function of $Y$: once we know the value of $Y$, this determines the value of $\mathbb{E}[X|Y]$. Thus, since $Y$ is a random variable, it follows that $\mathbb{E}[X|Y]$ is also a random variable. Furthermore, the distribution of $\mathbb{E}[X|Y]$ is fully determined by the distribution of $Y$:

$$\mathbb{P}(\mathbb{E}[X|Y] = \mathbb{E}[X|Y = y_j]) = \mathbb{P}(Y = y_j) = p_j^Y.$$
Example, cont’d: The possible values of the random variable $\mathbb{E}[X|S]$ are \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5, 6\} and the probability distribution of $\mathbb{E}[X|S]$ is

\[
\begin{align*}
\mathbb{P}(\mathbb{E}[X|S] = 1) &= \mathbb{P}(S = 2) = \frac{1}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 1.5) &= \mathbb{P}(S = 3) = \frac{2}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 2) &= \mathbb{P}(S = 4) = \frac{3}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 2.5) &= \mathbb{P}(S = 5) = \frac{4}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 3) &= \mathbb{P}(S = 6) = \frac{5}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 3.5) &= \mathbb{P}(S = 7) = \frac{6}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 4) &= \mathbb{P}(S = 8) = \frac{5}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 4.5) &= \mathbb{P}(S = 9) = \frac{4}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 5) &= \mathbb{P}(S = 10) = \frac{3}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 5.5) &= \mathbb{P}(S = 11) = \frac{2}{36} \\
\mathbb{P}(\mathbb{E}[X|S] = 6) &= \mathbb{P}(S = 12) = \frac{1}{36}
\end{align*}
\]
Since the conditional expectation $E[X|Y]$ is a random variable, we can calculate its expectation, which will just be a real number. In fact, our next result, which is known as the **law of the iterated expectation**, asserts that this is just the ordinary expectation of $X$.

**Theorem**

Let $X$ and $Y$ be random variables with values in the sets $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$, respectively, and joint distribution $p_{ij} = P(X = x_i, Y = y_j)$. Then

$$E\{E[X|Y]\} = E[X].$$

**Remark:** The result is important for much the same reason that conditional probabilities are important. When we are unable to calculate $E[X]$ directly, perhaps because the distribution of $X$ is complicated, we may be able to calculate $E[X]$ indirectly, by first calculating all of the conditional expectations of $X$ given $Y = y_j$ for the different possible values of $Y$ and then averaging these conditional expectations using the distribution of $Y$. This illustrates a recurrent theme in probability: when you can't solve a problem directly, try to solve it by conditioning on some other information and then averaging the conditional solutions.
Proof: By the definition of the expectation of a random variable and the definition of the conditional expectation \( \mathbb{E}[X|Y] \), we know that

\[
\mathbb{E} \{\mathbb{E}[X|Y]\} = \sum_{j=1}^{m} p_j^Y \cdot \mathbb{E}[X|Y = y_j]
\]

\[
= \sum_{j=1}^{m} p_j^Y \cdot \frac{1}{p_j^Y} \sum_{i=1}^{n} p_{i,j} \cdot x_i
\]

\[
= \sum_{j=1}^{m} \sum_{i=1}^{n} p_{i,j} \cdot x_i
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} p_{i,j} \cdot x_i
\]

\[
= \sum_{i=1}^{n} x_i \cdot \sum_{j=1}^{m} p_{i,j}
\]

\[
= \sum_{i=1}^{n} x_i \cdot p_i^X
\]

\[
= \mathbb{E}[X].
\]
Although the joint distribution of two random variables always determines the marginal distributions, the converse is not true in general: simply knowing the marginal distributions of two variables does not provide us with sufficient information to reconstruct the joint distribution. However, an important exception to this rule holds when the variables are independent of one another.

**Definition**

1. Let $X$ and $Y$ be random variables with values in the sets $E$ and $F$, respectively. Then $X$ and $Y$ are said to be **independent** if for all subsets $A \subset E$ and $B \subset F$, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent, i.e.,

$$
\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).
$$

2. Similarly, the random variables $X_1, \cdots, X_n$ with values in the sets $E_1, \cdots, E_n$, respectively, are said to be **independent** if for all subset $A_1 \subset E_1, \cdots, A_n \subset E_n$, the events $\{X_i \in A_i\}$ are independent, i.e.,

$$
\mathbb{P}(X_1 \in A_1, \cdots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i).
$$
In other words, two random variables $X$ and $Y$ are independent if knowing the value of $X$ provides us with no information concerning the value of $Y$. Let us investigate what independence tells us about the joint and marginal distributions of two random variables.

Suppose that $X$ and $Y$ are independent random variables with values in the sets
\[ \{x_1, \cdots, x_n\} \text{ and } \{y_1, \cdots, y_m\}, \]
respectively. Then

\[
p_{i,j} = \Pr(X = x_i, Y = y_j) \\
= \Pr(X \in \{x_i\}, Y \in \{y_j\}) \\
= \Pr(X \in \{x_i\}) \cdot \Pr(Y \in \{y_j\}) \\
= \Pr(X = x_i) \cdot \Pr(Y = y_j) \\
= p_i^X p_j^Y
\]

which shows that the joint distribution of two independent random variables is equal to the product of their marginal distributions. In fact, this condition also implies independence.
Another consequence of independence is that the conditional distribution of $X$ given $Y = y_j$ does not depend on $y_j$. This follows from a simple calculation:

$$
P(X = x_i | Y = y_j) = p_{i|j} = \frac{p_{i,j}}{p_j^Y} = \frac{p_i^X \cdot p_j^Y}{p_j^Y} = p_i^X = P(X = x_i).
$$

Similarly, if $X$ and $Y$ are independent, then the conditional expectation of $X$ given $Y = y_j$ is equal to the ordinary expectation of $X$, i.e., it doesn’t depend on the value of $Y$:

$$
\mathbb{E}[X | Y = y_j] = \sum_{i=1}^{n} p_{i|j} \cdot x_i = \sum_{i=1}^{n} p_i^X \cdot x_i = \mathbb{E}[X].
$$

In turn, if $X$ and $Y$ are independent, then we also have

$$
\mathbb{E}[X | Y] = \mathbb{E}[X],
$$

showing that $\mathbb{E}[X | Y]$ is not random in this case, i.e., it is equal to the constant $\mathbb{E}[X]$. 

Jay Taylor (ASU)
One of the most important consequences of independence is described in the next theorem.

**Theorem**

Suppose that $X$ and $Y$ are independent random variables with values in the sets $E = \{x_1, \cdots, x_n\}$ and $F = \{y_1, \cdots, y_m\}$, respectively. If $f : E \rightarrow \mathbb{R}$ and $g : F \rightarrow \mathbb{R}$ are real-valued random variables, then

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

In particular, by taking $f(x) = x$ and $g(y) = y$, we have

$$
\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]
$$

In words, the expected value of the product of two independent random variables is equal to the product of the expected values of each of the random variables. Be aware that this result does not hold in general if the random variables are not independent.
Proof: Using the law of the unconscious statistician, we can write the expected value of the product \( f(X)g(Y) \) as

\[
\mathbb{E}[f(X)g(y)] = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{i,j}f(x_i)g(y_j)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} p_{i}^{X} p_{j}^{Y} f(x_i)g(y_j)
\]

\[
= \sum_{i=1}^{n} p_{i}^{X} f(x_i) \sum_{j=1}^{m} p_{j}^{Y} g(y_j)
\]

\[
= \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)].
\]
Finally, we close this section by defining conditional independence of two random variables.

**Definition**

Suppose that $W$ is a random variable that takes values in the set $\{w_1, \cdots, w_p\}$. Two random variables $X$ and $Y$ with values in the sets $E$ and $F$, respectively, are said to be conditionally independent given $W$ if for each value $w_i$ and any pair of subsets $A \subset E$ and $B \subset F$, the events $X \in A$ and $Y \in B$ are conditionally independent given $W = w_i$, i.e.,

$$
P(X \in A, Y \in B | W = w_i) = P(X \in A | W = w_i)P(Y \in B | W = w_i).
$$

Informally, two random variables $X$ and $Y$ are conditionally independent given $W$ if all of the dependence between $X$ and $Y$ is determined by $W$, i.e., $X$ and $Y$ covary solely because they both depend on a common variable, $W$. 

**Problem:** Suppose that a coin with probability $p$ of landing on heads is tossed $n$ times. If the tosses are independent of each other, what is the probability that exactly $k$ of the tosses land on heads?

**Solution:** Let $X_i \in \{H, T\}$ denote the outcome of the $i$'th toss and let $X$ denote the total number of tosses that land on heads. There are $2^n$ possible outcomes for this experiment, each of which can be represented by a sequence of $n$ characters drawn from the set $\{H, T\}$. Furthermore, the event that we are interested in consists of those sequences that contain exactly $k$ occurrences of the letter $H$ and $n - k$ occurrences of the letter $T$. Let us denote this event by the set $D_{k,n}$. Since each element in $D_{k,n}$ can be identified by specifying the $k$ tosses that landed on heads, we see that $D_{k,n}$ contains

$$|D_{k,n}| = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

possible outcomes.
Let us investigate the probabilities of the different outcomes in $D_{k,n}$. To be concrete, suppose that $n = 4$ and $k = 2$, in which case there are $\binom{4}{2} = 6$ outcomes in $D_{2,4}$. Because the tosses are independent, the probability of each outcome is equal to the product of the probabilities of each of the tosses, giving the following result:

\[
\begin{align*}
\mathbb{P}(X_1 = H, X_2 = H, X_3 = T, X_4 = T) &= \mathbb{P}(X_1 = H)\mathbb{P}(X_2 = H)\mathbb{P}(X_3 = T)\mathbb{P}(X_4 = T) \\
&= p^2(1 - p)^2; \\
\mathbb{P}(X_1 = H, X_2 = T, X_3 = H, X_4 = T) &= p(1 - p)p(1 - p) = p^2(1 - p)^2 \\
\mathbb{P}(X_1 = H, X_2 = T, X_3 = T, X_4 = H) &= p(1 - p)^2 p = p^2(1 - p)^2 \\
\mathbb{P}(X_1 = T, X_2 = H, X_3 = H, X_4 = T) &= (1 - p)p^2(1 - p) = p^2(1 - p)^2 \\
\mathbb{P}(X_1 = T, X_2 = H, X_3 = T, X_4 = H) &= (1 - p)p(1 - p)p = p^2(1 - p)^2 \\
\mathbb{P}(X_1 = T, X_2 = T, X_3 = H, X_4 = H) &= (1 - p)^2 p^2 = p^2(1 - p)^2.
\end{align*}
\]

This shows that each outcome in $D_{2,4}$ has probability $p^2(1 - p)^2$ and so

\[
\mathbb{P}(X = 2) = \mathbb{P}(D_{2,4}) = \binom{4}{2} p^2(1 - p)^2.
\]
In general, because the tosses are independent, the probability of a particular outcome depends only on the numbers of tosses that land on heads and on tails and not on the order in which these occur. In particular, the probability of each outcome in $D_{k,n}$ is just $p^k(1 - p)^{n-k}$ since $k$ of the tosses must land on heads and $n - k$ must land on tails. For example, the probability of the outcome in which the first $k$ tosses land on heads and the last $n - k$ tosses land on tails is

$$P(X_1 = \cdots = X_k = H, X_{k+1} = \cdots = X_n = T) =$$

$$= P(X_1 = H) \cdots P(X_k = H) \cdot P(X_{k+1} = T) \cdots P(X_n = T)$$

$$= p^k(1 - p)^{n-k}.$$ 

Then, since $D_{k,n}$ contains $\binom{n}{k}$ outcomes, it follows that

$$P(X = k) = P(D_{k,n}) = \binom{n}{k} p^k(1 - p)^{n-k}.$$
The distribution derived on this preceding slides is sufficiently important to have its own name.

**Definition**

A random variable $X$ is said to have the **binomial distribution** with parameters $n \geq 1$ and $p \in [0, 1]$ if $X$ takes values in the set $\{0, 1, \cdots, n\}$ with probabilities

$$
P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, \cdots, n.
$$

In this case, we also say that $X$ is a **binomial random variable** and we can write $X \sim$ Binomial$(n, p)$. If $n = 1$, then $X$ is said to be a Bernoulli random variable with success probability $p$.

**Remark:** Notice that there is not just one binomial distribution, but rather an entire family of such distributions depending on the two parameters $n$ and $p$. 
To confirm that the binomial distribution is a proper probability distribution on \{0, 1, \ldots, n\}, we need to check that the sum of the probabilities of these outcomes is equal to 1. In fact,

\[
\sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} = (p + 1 - p)^n = 1,
\]

where the first identity is a consequence of the binomial theorem:

\[
(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.
\]
The binomial distribution commonly occurs in two settings:

- Suppose that a series of $n$ independent trials is performed and that each trial has probability $p$ of resulting in a ‘success’ (however we may define that). Then the total number of successes has the binomial distribution with parameters $n$ and $p$.

- Suppose that we sample $n$ times with replacement from a population containing two types of individuals with frequencies $p$ and $1 - p$. Then the number of individuals in our sample belonging to the first type is binomially distributed with parameters $n$ and $p$.

**Example:** Every indicator variable $I_A$ is a Bernoulli random variable with parameter $p = P(A)$. 
**Problem:** Suppose that a coin is tossed $n$ times but now assume that each toss lands on heads with probability $p$, on tails with probability $q$, and on edge with probability $r$. Here $p + q + r = 1$, so that one of these three outcomes occurs on each toss. Assuming that the tosses are independent of one another, what is the probability of the event that $n_1$ of the tosses land on heads, $n_2$ land on tails, and $n_3$ land on edge?

**Solution:** Let $Y_i \in \{H, T, E\}$ denote the outcome of the $i$'th toss and let $X_1$ be the number of tosses landing on heads, let $X_2$ be the number landing on tails, and let $X_3$ be the number landing on edge. In this case, there are $3^n$ possible outcomes, each of which can be represented by a sequence of $n$ characters from the set $\{H, T, E\}$. Let $D_{n_1, n_2, n_3}$ be the set of sequences which contain $n_1$ occurrences of $H$, $n_2$ occurrences of $T$, and $n_3$ occurrences of $E$. Notice that each outcome in $D_{n_1, n_2, n_3}$ corresponds to a partition of a set of $n$ objects into three subsets $I_H$, $I_T$ and $I_E$ containing $n_1$, $n_2$, and $n_3$ objects, respectively.
For example, if \( n_1 = 2, \ n_2 = 1 \) and \( n_3 = 1 \), then \( D_{2,1,1} \) contains the following sequences:

\[
\begin{align*}
HHTE & \iff I_H = \{1, 2\}, \ I_T = \{3\}, \ I_E = \{4\} \\
HHET & \iff I_H = \{1, 2\}, \ I_T = \{4\}, \ I_E = \{3\} \\
HTHE & \iff I_H = \{1, 3\}, \ I_T = \{2\}, \ I_E = \{4\} \\
HEHT & \iff I_H = \{1, 3\}, \ I_T = \{4\}, \ I_E = \{2\} \\
HTEH & \iff I_H = \{1, 3\}, \ I_T = \{2\}, \ I_E = \{4\} \\
HETH & \iff I_H = \{1, 3\}, \ I_T = \{2\}, \ I_E = \{4\} \\
THHE & \iff I_H = \{2, 3\}, \ I_T = \{1\}, \ I_E = \{4\} \\
EHHT & \iff I_H = \{2, 3\}, \ I_T = \{4\}, \ I_E = \{1\} \\
THEH & \iff I_H = \{2, 4\}, \ I_T = \{1\}, \ I_E = \{3\} \\
EHTH & \iff I_H = \{2, 4\}, \ I_T = \{3\}, \ I_E = \{1\} \\
TEHH & \iff I_H = \{3, 4\}, \ I_T = \{1\}, \ I_E = \{2\} \\
ETHH & \iff I_H = \{3, 4\}, \ I_T = \{2\}, \ I_E = \{1\}
\end{align*}
\]
In general, the number of ways to partition a set of \( n \) elements into \( k \) subsets containing \( n_1, n_2, \cdots, n_k \) elements each is given by the **multinomial coefficient**

\[
\binom{n}{n_1, n_2, \cdots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}.
\]

Thus \( D_{n_1, n_2, n_3} \) contains \( \binom{n}{n_1, n_2, n_3} \) outcomes and indeed

\[
|D_{2,1,1}| = \binom{4}{2,1,1} = \frac{4!}{2!1!1!} = 12
\]

as can also be seen by the direct enumeration of the outcomes in this set on the preceding slide.
As in the original problem, the probability of any particular outcome only depends on the numbers of tosses landing on heads, tails or edge, respectively, not on the order in which they occur. For example, in the specific case $n_1 = 2, n_2 = n_3 = 1$, we have

\[
\mathbb{P}(HHTE) = \mathbb{P}(Y_1 = H, Y_2 = H, Y_3 = T, Y_4 = E) \\
= \mathbb{P}(Y_1 = H)\mathbb{P}(Y_2 = H)\mathbb{P}(Y_3 = T)\mathbb{P}(Y_4 = E) \\
= p^2 qr,
\]

and indeed all twelve outcomes in $D_{2,1,1}$ have this same probability, giving

\[
\mathbb{P}(X_1 = 2, X_2 = 1, X_3 = 1) = \binom{4}{2,1,1} p^2 qr.
\]

In the general case, the probability of any particular outcome landing on $n_1$ heads, $n_2$ tails and $n_3$ edges is $p^{n_1} q^{n_2} r^{n_3}$ and so

\[
\mathbb{P}(X_1 = n_1, X_2 = n_2, X_3 = n_3) = \binom{n}{n_1, n_2, n_3} p^{n_1} q^{n_2} r^{n_3}.
\]
Like the binomial distribution, this distribution is important enough to merit its own name.

**Definition**

A set of random variables \((X_1, \cdots, X_k)\) is said to have the **multinomial distribution** with parameters \(n\) and \(p = (p_1, \cdots, p_k)\) if \((X_1, \cdots, X_k)\) takes values in the set

\[
\{(n_1, \cdots, n_k) : n_1, \cdots, n_k \in \mathbb{N}, n_1 + \cdots + n_k = n\}
\]

with probabilities

\[
P(X_1 = n_1, \cdots, X_k = n_k) = \binom{n}{n_1, \cdots, n_k} p_1^{n_1} \cdots p_k^{n_k}.
\]
The multinomial distribution typically occurs in scenarios where we need to consider more than just two possible outcomes or types of individuals.

- Suppose that a series of $n$ independent trials is performed and that each trial can result in any one of $k$ outcomes. If the probability of the $i$'th outcome is $p_i$, then the numbers of trials resulting in each kind of outcome are multinomially distributed with parameters $n$ and $(p_1, \cdots, p_k)$.

- Suppose that we sample $n$ times with replacement from a population containing $k$ types of individuals with frequencies $p_1, \cdots, p_k$. Then the numbers of individuals in our sample of each type are multinomially distributed with parameters $n$ and $(p_1, \cdots, p_k)$. 
Exercise: Suppose that an urn contains five red balls, three blue balls, and two yellow balls. If five balls are sampled with replacement from this urn, calculate the probability that the sample will contain two red balls, one blue ball and two yellow balls.
**Exercise:** Suppose that an urn contains five red balls, three blue balls, and two yellow balls. If five balls are sampled with replacement from this urn, calculate the probability that the sample will contain two red balls, one blue ball and two yellow balls.

**Solution:** Using the multinomial distribution with $n = 5$, $p_1 = 0.5$, $p_2 = 0.3$ and $p_3 = 0.2$, we have

\[
\mathbb{P}(X_1 = 2, X_2 = 1, X_3 = 2) = \binom{5}{2, 1, 2} p_1^2 p_2 p_3^2
\]

\[
= \frac{5!}{2!1!2!} 0.5^2 \cdot 0.3 \cdot 0.2^2
\]

\[
= 30 \cdot 0.003
\]

\[
= 0.09.
\]
Sampling Distributions

Both the binomial and the multinomial distribution arise in problems involving sampling with replacement. However, for certain kinds of applications it may be more natural to sample without replacement. Before considering the general formulation of this problem, let us examine a special case.

**Problem:** Suppose that an urn contains five red balls, three blue balls, and two yellow balls. If five balls are sampled without replacement from this urn, calculate the probability that the sample will contain two red balls, one blue ball and two yellow balls.

**Solution:** If we imagine labeling the balls 1 through 10, then each outcome can be described by a choice of 5 numbers from the set \{1, \ldots, 10\}, which shows that there are \( \binom{10}{5} = 252 \) possible outcomes. Furthermore, since each of these outcomes is equally likely, it follows that the probability of any one of them is 1/252.
To find the probability that our sample contains two red balls, one blue ball and two yellow balls, let us count the number of ways in which this event could occur. Since there are five red balls, there are \( \binom{5}{2} = 10 \) ways to choose two these for our sample. Similarly, since there are three blue balls, there are clearly \( \binom{3}{1} = 3 \) ways to choose one of these for our sample. Finally, since we are sampling without replacement, there is just one way to choose two yellow balls since the urn itself only contains two yellow balls.

Notice that the choices of the balls of each of the three colors are independent of one another. For example, whether we choose balls \{1, 2\} or \{2, 5\} from the red balls has no bearing on whether we choose ball 6 or ball 7 or ball 8 as the lone blue ball in the sample. (Here I am assuming that the red balls are labeled 1, \cdots, 5, the blue balls are labeled 6, 7, 8 and the yellow balls are labeled 9, 10.) Thus to find the total number of choices, we need to multiply the numbers of choices of the balls of each color:

\[
\binom{5}{2} \times \binom{3}{1} \times \binom{2}{2} = 10 \times 3 \times 1 = 30.
\]
Finally, since the probability of each possible outcome is $1/252$, it follows that the probability of sampling two blue balls, one red ball, and two yellow balls without replacement is

$$\binom{5}{2} \binom{3}{1} \binom{2}{2} = \frac{30}{252} = \frac{5}{42}.$$ 

**General problem:** Now suppose that a population is divided into $K$ classes of individuals, such that the first class contains $m_1$ individuals, the second class contains $m_2$ individuals, and so forth. Let $m = m_1 + \cdots + m_K$ denote the total number of individuals in the population. If $n$ individuals are sampled without replacement from this population, what is the probability that the sample contains $n_1$ individuals from the first class, $n_2$ individuals from the second class, and so forth?
Solution: If the individuals in the population are labeled 1, ⋯ , m, then the sample space contains \( \binom{m}{n} \) possible outcomes, one corresponding to each group of \( n \) individuals that might be sampled. (Notice that the order in which we sample these individuals does not matter.) Furthermore, since each of these outcomes is equally likely - we don’t selectively choose certain individuals over others - the probability of each outcome is \( 1/\binom{m}{n} \).

To calculate the probability of the event in which our sample contains \( n_1 \) individuals from the first class, \( n_2 \) from the second class, and so forth, we need to count the number of outcomes that satisfy these conditions. However, there are \( \binom{m_1}{n_1} \) ways of choosing \( n_1 \) individuals from the first class, \( \binom{m_2}{n_2} \) ways of choosing \( n_2 \) individuals from the second class, etc., and each of these choices is independent of the others. Thus, if we write \( X_i \) for the number of individuals from the \( i \)'th class that are included in the sample, then

\[
P(X_1 = n_1, \cdots, X_k = n_k) = \frac{\binom{m_1}{n_1} \cdots \binom{m_k}{n_k}}{\binom{m}{n}}.
\]

Remark: When \( K = 2 \), this distribution is known as the **hypergeometric distribution**.
Variance

On its own, the expectation of a random variable tells us relatively little about its distribution: it may be the case that the difference between the variable and its mean is always fairly small, but this need not be true in general. For this reason, it is useful to have a number that quantifies the spread or the dispersion of a random variable around its mean. One such metric is called the variance, which is defined below.

**Definition**

Suppose that $X$ is a random variable with expectation $\mu = \mathbb{E}[X]$. The variance of $X$ is the quantity

$$Var(X) \equiv \mathbb{E}[(X - \mu)^2],$$

while the standard deviation of $X$ is the square root of the variance

$$SD(X) = \sqrt{Var(X)}.$$
If $X$ takes values in the set $E = \{x_1, \cdots, x_n\} \subset \mathbb{R}$ with probabilities $p_1, \cdots, p_n$, then the variance of $X$ can be written as

$$Var(X) = \sum_{i=1}^{n} p_i (x_i - \mu)^2.$$ 

This shows that the variance is always non-negative, $Var(X) \geq 0$, since the right-hand side is a sum of non-negative quantities. Furthermore, $Var(X) = 0$ if and only if $\mathbb{P}(X = \mu) = 1$. In other words, if the variance of a random variable is equal to 0, then that variable is essentially a constant.

In general, the larger the variance of a random variable $X$, the more dispersed that variable will be around its mean. If $Var(X)$ is small, then $X$ will typically take values close to $\mu$. However, if $Var(X)$ is very large, then $X$ may take values that are very different from the mean.
There is an alternative formula that can be used to calculate the mean. Letting \( \mu = \mathbb{E}[X] \) and using the linearity of expectations, we have

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

In words, the variance is equal to the expectation of the square of a random variable minus the square of the expectation.

**Example:** Let \( I_A \) be an indicator variable with \( p = \mathbb{E}[I_A] = \mathbb{P}(A) \). Then \( I_A^2 = I_A \) and so the variance of \( I_A \) is

\[
\text{Var}(I_A) = \mathbb{E}[I_A^2] - \mathbb{E}[I_A]^2 = \mathbb{E}[I_A] - \mathbb{E}[I_A]^2 = p - p^2 = p(1 - p).
\]
Exercise: Let $X$ be the number that turns up when a fair six-sided die is rolled. Find the variance and the standard deviation of $X$. 

Solution: First calculate the expectation of $X$: 

$$
\mu = \mathbb{E}[X] = \sum_{i=1}^{6} \frac{1}{6} \cdot i = \frac{21}{2}.
$$

Then the variance of $X$ is 

$$
\text{Var}(X) = \mathbb{E}(\hat{X}^2) - \mu^2 = \sum_{i=1}^{6} \frac{1}{6} \cdot i^2 - \left(\frac{21}{2}\right)^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.
$$

while the standard deviation is 

$$
\text{SD}(X) = \sqrt{\frac{35}{12}} \approx 1.7078.
$$
**Exercise:** Let $X$ be the number that turns up when a fair six-sided die is rolled. Find the variance and the standard deviation of $X$.

**Solution:** First calculate the expectation of $X$:

$$
\mu = \mathbb{E}[X] = \sum_{i=1}^{6} \frac{1}{6} \cdot i = \frac{7}{2}.
$$

Then the variance of $X$ is

$$
\text{Var}(X) = \mathbb{E}[X^2] - \mu^2 = \sum_{i=1}^{6} \frac{1}{6} \cdot i^2 - \frac{49}{4} = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}
$$

while the standard deviation is

$$
\text{SD}(X) = \sqrt{\frac{35}{12}} \approx 1.7078.
$$
Standard Deviation versus Variance

The main difference between the variance and the standard deviation is in the units in which they are expressed.

- The expectation of $X$ has the same units as $X$.
- The standard deviation of $X$ has the same units as $X$.
- The variance of $X$ has the units of $X^2$.

For example, if $X$ is the height of a randomly sampled individual measured in cm, then the mean and the standard deviation of $X$ will both be expressed in units of cm, while the variance will be expressed in units of cm$^2$. 
The next theorem shows how the variance of a random variable changes under linear transformation.

**Theorem**

*Suppose that $X$ is a random variable with expectation $\mu_X = \mathbb{E}[X]$ and let $Y = kX + b$, where $k$ and $b$ are constants. Then*

$$\text{Var}(Y) = k^2 \text{Var}(X)$$

**Proof:** Because expectations transform linearly, we know that

$$\mu_Y \equiv \mathbb{E}[Y] = \mathbb{E}[kX + b] = k\mu_X + b.$$

Consequently,

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

$$= \mathbb{E}[(kX + b - k\mu_X - b)^2]$$

$$= k^2 \mathbb{E}[(X - \mu_X)^2]$$

$$= k^2 \text{Var}(X).$$
Covariance

Whereas the variance quantifies the variability of a single random variable, the covariance measures the strength of the statistical association between two different random variables.

**Definition**

Suppose that \( X \) and \( Y \) are random variables with means \( \mu_X = \mathbb{E}[X] \) and \( \mu_Y = \mathbb{E}[Y] \). Then the covariance of \( X \) and \( Y \) is the quantity

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

Also, if the standard deviations of \( X \) and \( Y \) are positive, say \( \sigma_X = SD(X) > 0 \) and \( \sigma_Y = SD(Y) > 0 \), then the correlation between \( X \) and \( Y \) is the quantity

\[
\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.
\]
Like the variance, there is an alternative expression for the covariance that is often more convenient in calculations.

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

\[
= \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y]
\]

\[
= \mathbb{E}[XY] - \mu_X \mathbb{E}[Y] - \mu_Y \mathbb{E}[X] + \mu_X \mu_Y
\]

\[
= \mathbb{E}[XY] - 2\mu_X \mu_Y + \mu_X \mu_Y
\]

\[
= \mathbb{E}[XY] - \mu_X \mu_Y.
\]

Thus the covariance between two random variables is equal to the difference between the expected value of the product of the variables and the product of the expected values of the variables.
The covariance of two random variables is a measure of the linear dependence between those two variables.

- The covariance will be positive if whenever one variable is unusually large, so is the other.
- The covariance will be negative if whenever one variable is unusually large, the other is unusually small.
- If $X$ and $Y$ are independent, then

\[
\text{Cov}(X, Y) = \mathbb{E}[XY] - \mu_X \mu_Y = \mathbb{E}[X] \mathbb{E}[Y] - \mu_X \mu_Y = 0.
\]

Unfortunately, the converse of this last result is not true: the covariance between two variables may vanish even when those variables are not independent. Indeed, we say that two random variables $X$ and $Y$ are \textbf{uncorrelated} if $\text{Cov}(X, Y) = 0$. 
**Example:** Suppose that $X$ is a random variable with distribution

$$\mathbb{P}(X = 0) = \mathbb{P}(X = 1) = \mathbb{P}(X = -1) = \frac{1}{3},$$

and define the random variable $Y$ by

$$Y = \begin{cases} 
0 & \text{if } X \neq 0 \\
1 & \text{if } X = 0.
\end{cases}$$

Show that $X$ and $Y$ are uncorrelated, but not independent.
Example: Suppose that $X$ is a random variable with distribution
\[ P(X = 0) = P(X = 1) = P(X = -1) = \frac{1}{3}, \]
and define the random variable $Y$ by
\[ Y = \begin{cases} 0 & \text{if } X \neq 0 \\ 1 & \text{if } X = 0. \end{cases} \]
Show that $X$ and $Y$ are uncorrelated, but not independent.

Solution: Since $XY = 0$, it follows that $E[XY] = 0$. Also, $\mu_X = E[X] = 0$ and so
\[ \text{Cov}(X, Y) = E[XY] - \mu_X \mu_Y = 0, \]
which shows that $X$ and $Y$ are uncorrelated. On the other hand, $X$ and $Y$ are not independent since
\[ P(X = 0 | Y = 1) = 1 \neq \frac{1}{3} = P(X = 0). \]
Our next theorem lists several useful properties of covariances.

**Theorem**

(a) $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

(b) $\text{Cov}(X, X) = \text{Var}(X)$

(c) $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$ for any constants $a, b, c, d$;

(d) $\text{Cov} \left( \sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$.

**Proof:** For (a), observe that

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

For (b), observe that

$$\text{Cov}(X, X) = \mathbb{E}[(X - \mu_X)(X - \mu_X)] = \mathbb{E}[(X - \mu_X)^2] = \text{Var}(X).$$
To prove (c), observe that $\mathbb{E}[aX + b] = a\mu_X + b$ and $\mathbb{E}[cY + d] = c\mu_Y + d$. Consequently,

\[
\text{Cov}(aX + b, cY + d) = \mathbb{E}[(aX + b - a\mu_X - b)(cY + d - c\mu_Y - d)] \\
= ac\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
= ac \text{Cov}(X, Y).
\]

Finally, to prove (d), observe that

\[
\mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \mu_{X_i} \\
\mathbb{E} \left[ \sum_{j=1}^{m} Y_j \right] = \sum_{j=1}^{m} \mathbb{E}[Y_j] = \sum_{j=1}^{m} \mu_{Y_j}.
\]
Substituting these expressions into the definition of the covariance gives

\[
\text{Cov} \left( \sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j \right) = \mathbb{E} \left[ \sum_{i=1}^{n} (X_i - \mu_{X_i}) \sum_{j=1}^{m} (Y_j - \mu_{Y_j}) \right]
\]

\[
= \mathbb{E} \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}) \right]
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbb{E} \left[ (X_i - \mu_{X_i})(Y_j - \mu_{Y_j}) \right]
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j).
\]
Part (d) of this last theorem has several useful consequences. First, given any set of random variables $X_1, \cdots, X_n$, we have

\[
\text{Var} \left( \sum_{i=1}^{n} X_i \right) = \text{Cov} \left( \sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j \right)
\]

\[
= \sum_{i=1}^{n} \sum_{j=1}^{n} \text{Cov}(X_i, X_j)
\]

\[
= \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)
\]

\[
= \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j).
\]

In other words, the variance of a sum of random variables is equal to the sum of the variances plus the sum of the covariances.
In particular, if $X_1, \cdots, X_n$ are independent, then

$$
\text{Var} \left( \sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{1 \leq i < j \leq n} \text{Cov}(X_i, X_j) = \sum_{i=1}^{n} \text{Var}(X_i)
$$

since independence implies that all of the covariances $\text{Cov}(X_i, X_j), j \neq i$ are zero. Thus the variance of a sum of independent random variables is equal to the sum of the variances. This result is one that is used often.
Mean and Variance of the Binomial Distribution

We can use these results to calculate the mean and the variance of the binomial distribution with parameters $n$ and $p$. Suppose that we perform a sequence of $n$ independent trials and that the $i$’th trial has probability $p$ of resulting in a success. Let $X_i$ be the indicator variable for the event that the $i$’th trial is a success and let

$$X = X_1 + \cdots + X_n$$

be the total number of successes. Then $X \sim \text{Binomial}(n, p)$.

We first calculate the expectation of $X$. By linearity of expectations, this is

$$E[X] = E \left[ \sum_{i=1}^{n} X_i \right] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} p = np.$$ 

Thus the expected number of successes is equal to the number of trials multiplied by the probability that any one trial is successful.
To calculate the variance, we use the result on the variance of a sum of independent variables:

\[
\text{Var}(X) = \text{Var} \left( \sum_{i=1}^{n} X_i \right) \\
= \sum_{i=1}^{n} \text{Var}(X_i) \\
= np(1 - p),
\]

since the variance of each of the indicator variables $X_i$ is just $p(1 - p)$. In particular, notice that the variance is 0 whenever $p = 0$ or $p = 1$, and the variance is maximized when $p = 1/2$. 
We next turn our attention to the correlation $\rho(X, Y)$. To see why this quantity is useful, consider the correlation between the linear transformations of two variables, say $aX + b$ and $cY + d$, where $a > 0$ and $c > 0$. We first observe that

$$SD(aX + b) = aSD(X) \text{ and } SD(cY + d) = cSD(Y).$$

Consequently,

$$\rho(aX + b, cY + d) = \frac{Cov(aX + b, cY + d)}{SD(aX + b)SD(cY + d)}$$

$$= \frac{ac \cdot Cov(X, Y)}{aSD(X) \cdot cSD(Y)}$$

$$= \frac{Cov(X, Y)}{SD(X)SD(Y)}$$

$$= \rho(X, Y),$$

which shows that the correlation does not change under linear transformations. For this reason, we say that the correlation is scale- and shift-invariant, which in practice means that it does not depend on the units in which $X$ and $Y$ are measured, i.e., the correlation is a **dimensionless** quantity.
Another important property of correlations is that:

**Theorem**

For any pair of random variables $X$ and $Y$ with positive standard deviations,

$$-1 \leq \rho(X, Y) \leq 1.$$ 

**Proof:** Let $\sigma_X = SD(X)$ and $\sigma_Y = SD(Y)$. Then, since the variance of any random variable is non-negative, we have

$$0 \leq Var\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$$

$$= \frac{Var(X)}{\sigma_X^2} + \frac{Var(Y)}{\sigma_Y^2} + 2 \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$= 2(1 + \rho(X, Y)),$$

which shows that $\rho(X, Y) \geq -1$. A similar calculation starting with $Var\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right)$ shows that $\rho(X, Y) \leq 1$. 

□
It can be shown that if $\rho(X, Y) = 1$, then the following identity holds with probability 1

$$Y = \mu_Y + \frac{\sigma_Y}{\sigma_X} (X - \mu_X),$$

while if $\rho(X, Y) = -1$, then

$$Y = \mu_Y - \frac{\sigma_Y}{\sigma_X} (X - \mu_X).$$

holds with probability 1. In general, the closer $\rho(X, Y)$ is to $\pm 1$, the closer $X$ and $Y$ are to a pair of linearly related variables.
Example: Suppose that $X$ and $Y$ are independent binomially distributed random variables with parameters $n$ and $p$, and let $Z = X + Y$. Then $Z$ is a binomial random variable with parameters $2n$ and $p$ and $\text{Var}(X) = np(1 - p)$, $\text{Var}(Z) = 2np(1 - p)$ and

\[
\text{Cov}(X, Z) = \text{Cov}(X, X + Y) = \text{Cov}(X, X) + \text{Cov}(X, Y) = \text{Var}(X) = np(1 - p)
\]

\[
\rho(X, Z) = \frac{np(1 - p)}{\sqrt{np(1 - p) \cdot 2np(1 - p)}} = \frac{1}{\sqrt{2}}.
\]
Estimation of Means and Variances

Suppose that $X_1, \cdots, X_n$ are independent, identically-distributed (abbreviated i.i.d.) random variables, each with mean $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \text{Var}(X_i)$. We will think of these variables as arising from a series of $n$ independent experiments where our goal is to estimate $\mu$ and $\sigma^2$. Thus $X_i$ is the measurement taken on the $i$'th trial.

Let us define the sample mean $\bar{X}$ and the sample variance $V$ of this data by the following expressions

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

$$V = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

These quantities are often used as estimates of the mean $\mu$ and the variance $\sigma^2$. 

We first observe that

\[ E \left[ \bar{X} \right] = E \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu. \]

In other words, the expected value of the sample mean is equal to the true mean. In general, any estimator that has the property that its expectation is equal to the value of the unknown quantity that it is intended to estimate is said to be unbiased. Thus the sample mean is an unbiased estimator of the true mean.
Although unbiasedness is a desirable property of an estimator, we also care about the variance of the estimator about its expectation. For the sample mean, we can calculate

\[
\text{Var}(\bar{X}) = \left( \frac{1}{n} \right)^2 \text{Var} \left( \sum_{i=1}^{n} X_i \right) \\
= \left( \frac{1}{n} \right)^2 \sum_{i=1}^{n} \text{Var}(X_i) \\
= \frac{\sigma^2}{n},
\]

which shows that the variance of the sample mean is inversely proportional to the number of independent replicates that have been performed. Thus, as the number of trials increases, the variance of the sample mean decreases and the estimator becomes more accurate.
We can also show that the sample variance is an unbiased estimator of the true variance. We first observe that

\[ V = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu + \mu - \bar{X})^2 \]

\[ = \frac{1}{n-1} \left( \sum_{i=1}^{n} (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^{n} (X_i - \mu) + \sum_{i=1}^{n} (\bar{X} - \mu)^2 \right) \]

\[ = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu)^2 - \frac{n}{n-1} (\bar{X} - \mu)^2. \]

Taking expectations gives

\[ \mathbb{E}[V] = \frac{1}{n-1} \sum_{i=1}^{n} \mathbb{E}[(X_i - \mu)^2] - \frac{n}{n-1} \text{Var}(\bar{X}) \]

\[ = \frac{n}{n-1} \sigma^2 - \frac{n}{n-1} \sigma^2 \]

\[ = \sigma^2. \]
Informally, the reason that we must divide by $n - 1$ rather than $n$ to obtain an unbiased estimate of the variance is that the deviations between the $X_i$’s and the sample mean (which is computed using the observed data) tend to be smaller than the deviations between the $X_i$’s and the true mean. Indeed, it can be shown that

$$
\bar{X} = \arg \min_m \left( \sum_{i=1}^n (X_i - m)^2 \right),
$$

so that $\bar{X}$ minimizes the sum of the squared deviations between the data and any constant.
Tail Inequalities

Theorem

(Markov’s Inequality) Suppose that $X$ is a non-negative random variable, i.e., $X$ takes values in a set \( \{x_1, \cdots, x_n\} \subset [0, \infty) \), with mean \( \mu = \mathbb{E}[X] \). Then, for every real number \( M > 0 \),

\[
\mathbb{P}(X > M) \leq \frac{\mu}{M}.
\]

Proof: If we let \( p_i = \mathbb{P}(X = x_i) \), then

\[
\mu = \sum_{i=1}^{n} p_i \cdot x_i = \sum_{x_i \leq M} p_i \cdot x_i + \sum_{x_i > M} p_i \cdot x_i \\
\geq \sum_{x_i > M} p_i \cdot x_i \\
\geq \sum_{x_i > M} p_i \cdot M = M \sum_{x_i > M} p_i = M \cdot \mathbb{P}(X > M).
\]
**Theorem**

*(Chebyshev’s Inequality)* Suppose that $X$ is a random variable with mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \text{Var}(X)$. Then, for every real number $\epsilon > 0$,

$$\mathbb{P}(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$  

**Proof:** By applying Markov’s inequality to the non-negative random variable $Y = (X - \mu)^2$, we see that

$$\mathbb{P}(|X - \mu| > \epsilon) = \mathbb{P}(Y > \epsilon^2) \leq \frac{\mathbb{E}[Y]}{\epsilon^2} = \frac{\text{Var}(X)}{\epsilon^2}.$$  

\[\square\]
(Weak Law of Large Numbers) Suppose that $X_1, X_2, \cdots$ is a sequence of independent identically-distributed random variables, each having mean $\mu = \mathbb{E}[X_i]$ and variance $\sigma^2 = \text{Var}(X_i)$. Then, for any $\epsilon > 0$,

$$
\lim_{n \to \infty} \mathbb{P}\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right\} = 0
$$

Proof: For each $n \geq 1$, let $Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the sample mean of the first $n$ random variables and observe that

$$
\mathbb{E}[Z_n] = \mu \text{ and } \text{Var}(Z_n) = \frac{\sigma^2}{n}.
$$

By applying Chebyshev’s inequality to $Z_n$, we have

$$
\mathbb{P}( |Z_n - \mu| > \epsilon ) \leq \frac{\text{Var}(Z_n)}{\epsilon^2} = \frac{\sigma^2}{n \epsilon^2},
$$

which tends to 0 as $n \to \infty$ for any $\epsilon > 0$. \qed
**Example:** Suppose that we perform a sequence of independent trials, each of which could result in an event $A$ with probability $p = \mathbb{P}(A)$. Let $X_i$ be the indicator variable of the event that $A$ occurs on the $i$’th trial. Then

$$Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

is the frequency of the event $A$ in the first $n$ trials. Since $\mathbb{E}[X_i] = p$ and $\text{Var}(X_i) = p(1-p)$, it follows from the weak law of large numbers that

$$\mathbb{P}(|Z_n - p| > \epsilon) < \frac{p(1-p)}{n\epsilon^2}$$

which can be made arbitrarily small by taking $n$ sufficiently large. This shows that if an experiment is repeatable, i.e, if we can perform a series of independent trials that have the same distribution of outcomes, then we can estimate the probability of an event by its frequency in a large number of trials.
Exercise: Suppose that you are a quality control manager working for an aeronautical company and that you are investigating a new process for manufacturing a critical component in an airplane. As part of this investigation, you need to check that the weight of each manufactured component is within a certain margin of error of a prescribed value. To do so, you will weigh each component $n$ times and use the average of these $n$ independent measurements as an estimate for the true weight of the component. How large should $n$ be so that there is less than a 1% chance that your estimate of the mean will differ from the true mean by more than 1% of the true value?
Exercise: Suppose that you are a quality control manager working for an aeronautical company and that you are investigating a new process for manufacturing a critical component in an airplane. As part of this investigation, you need to check that the weight of each manufactured component is within a certain margin of error of a prescribed value. To do so, you will weigh each component \( n \) times and use the average of these \( n \) independent measurements as an estimate for the true weight of the component. How large should \( n \) be so that there is less than a 1% chance that your estimate of the mean will differ from the true mean by more than 1% of the true value?

Solution: This problem can be addressed in several ways. Let \( X_1, \cdots, X_n \) be the series of measurements and assume that the measurement process is unbiased, so that \( \mu = \mathbb{E}[X_i] \) is equal to the true weight of the component. Also, let \( Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i \) be the sample mean of the \( n \) measurements. Our aim is to find \( n \) so that

\[
P(|Z_n - \mu| > 0.01\mu) \leq 0.01.
\]
If we knew the value of $\sigma^2 = \text{Var}(X_i)$, then we could use the inequality

$$\mathbb{P}(|Z_n - \mu| > 0.01\mu) \leq \frac{\sigma^2}{n(0.01\mu)^2}$$

to deduce that we should take $n$ at least as large as

$$n \geq \frac{1}{0.01} \frac{\sigma^2}{(0.01\mu)^2} = \frac{\sigma^2}{10^6\mu^2}.$$ 

However, if (as seems more likely) we don’t know the value of $\sigma^2$, then it too will need to be estimated. This can be done by carrying out a preliminary series of $m$ measurements and using these to calculate the sample variance

$$V = \frac{1}{m-1} \sum_{i=1}^{m} (X_i - \bar{X})^2.$$
Then we could either substitute $V$ for $\sigma^2$ in the equation for $n$ given on the previous slide or, if we were especially risk averse, we could substitute an inflated value of $V$, say $V(1 + \kappa)$ with $\kappa > 0$, to guard against underestimates of $\sigma^2$:

$$n = \frac{V(1 + \kappa)}{10^6 \mu^2}.$$ 

An even more conservative approach is possible if we know that the measurements are bounded below by a value $m$ and above by a value $M$. In this case, the variance is certain to be bounded above by the quantity $(M - m)^2$ and so we can substitute this into the equation for $n$:

$$n = \frac{(M - m)^2}{10^6 \mu^2}.$$. 