Extension of the LTP (4) \(\text{Assuming all conditional probabilities are defined in what follows, if } C \text{ is in } \mathcal{F} \text{ then}
\)
\[
P(A \mid C) = \sum_{j=1}^{k} p(B_j \mid C) p(A \mid B_j \cap C).
\]

**Definition** Events \(A, B\) are independent if
\[
p(A \cap B) = p(A) \cdot p(B),
\]
which (as long as \(p(A) > 0, p(B) > 0\)) is equivalent to \(p(A \mid B) = p(A)\) \(\Rightarrow\) (Bayesian)
\[
\text{and } p(B \mid A) = p(B).
\]
Consequences of the definition of independence

1. If $A$ and $B$ are independent, then so are $A$ and $B^c$, $A^c$ and $B$, and $A^c$ and $B^c$.

2. Extension of the definition to more than 2 events:

Definition:

Given events $A_1, ..., A_k$, they are (mutually) independent if, for every subset $A_{i_1}, ..., A_{i_j}$ of $(A_1, ..., A_k)$ ($j = 2, ..., k$),

$$P(A_{i_1} \cap ... \cap A_{i_j}) = P(A_{i_1}) \cdot ... \cdot P(A_{i_j})$$
Interpretation of independence: $A$, $B$ independent $\iff$ information about $A$ doesn't change the chances associated with $B$, and vice versa.

Definition: Another extension of independence. Events $\{A_i, \ldots, A_k\}$ are conditionally independent given event $B$ if for every subset $\{A_{i_1}, \ldots, A_{i_j}\}$ of $\{A_i, \ldots, A_k\}$ ($j = 2, \ldots, k$)

$$P(A_{i_1} \cap \cdots \cap A_{i_j} \mid B) = \prod_{k=1}^j P(A_{i_k} \mid B).$$
Suppose that there is a machine that can take an ordinary coin and produce IID tosses of the coin with \( P(H) = \theta \), and \( \theta \) can be set to any value in \([0,1]\) with a dial on the machine's control panel. Someone sets the dial to a \( \theta \) value that's unknown to you and starts producing coin tosses \( T_1, T_2, ... \).

Suppose the first 10 tosses come out 1011100111. <i>Bits</i> (binary digits) \( HHTHHHTHHH \) (7 \( H \), 3 \( T \)).

(John Tukey)

Q: Is there information in these first 10 tosses that helps you to predict \( T_{11} \)?
A: Yes, definitely; it looks like \( \Theta \) is around \( \frac{2}{10} \), so you would predict \( E'' \) to be probabilistically
\[ E'' = H. \]
Now, suppose instead that you watched the person with the machine set the dial to \( \Theta = 0.81 \), so that \( \Theta \) is now known to you. The next 10 tosses came out
\[ H H A T H T H H H H \]
(8H, 2T). A: Is there information in these 10 tosses that helps you to predict the next toss?
A: No; you know that \( \Theta = 0.81 \), so there's no information in any of the \( E_p \).
That helps you to predict any of the other $Y_j$ given $\theta$, the $Y_i$ are independent. Thus the $Y_i$ are unconditionally dependent but conditionally independent given $\theta$.

(4 Apr 17)

Bayes's Theorem

Suppose that the events $B_1, \ldots, B_k$ partition the sample space in such a way that $P(B_j) > 0$ for all $j = 1, \ldots, k$. If $A$ is an event with $P(A) > 0$, then for each $i = 1, \ldots, k$

$$P(B_i | A) = \frac{P(B_i) P(A | B_i)}{P(A)}$$
and, by the LTP, this is

$$P(B_i | A) = \frac{P(B_i) \cdot P(A | B_i)}{\sum_{j=1}^k P(B_j) \cdot P(A | B_j)}.$$

How this theorem is used in Bayesian statistics: The $B_i$ represent unknown states of the world: they're all possible $- P(B_i) > 0$ — and only one of them is true, but you don't know which one. $A$ represents data: information that will help you identify the most probable $B_i$. 
Before the dataset $A$ arrives, you have background information about the plausibility of the $B_i$ that you can represent with prior probabilities $P(B_i)$. After the dataset $A$ arrives, you can use Bayes's Theorem to update your prior probabilities to posterior probabilities $P(B_i | A)$. The probabilities $P(A | B_i)$ represent how likely the dataset $A$ would be if $B_i$ were the actual unknown state; this is often called likelihood information.
(The denominator)

$P(A)$ does not depend on $B_i$, and can therefore be regarded as a normalizing constant. Put into Bayes's Theorem to make all the $P(B_i|A)$ add up to 1. Thus

$$P(B_i|A) = \frac{P(B_i) P(A|B_i)}{P(A)}$$

is interpreted as

$$(\text{posterior (information)}) = \frac{\text{prior (information)} \cdot \text{likelihood (information)}}{\text{(normalizing constant)}}.$$
**Random variables and their distributions**

**Example: Tay-Sachs Disease**

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>NNNNNN</td>
<td>T</td>
<td>N</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>TNNNNN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NTNNNN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNTNNT</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNNTNT</td>
<td>T</td>
<td>T</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>TNNNTN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNTNNT</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NTNTNN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNTNNN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNTNTN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNTNNN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>NNNNTN</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td>TTTTTT</td>
<td>N</td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
</tbody>
</table>

- | # of T-s babies = \( I \)
- **Definition**: Given a sample space \( S \) for an experiment \( E \), a (real-valued) random variable \( \text{rv} \) is a function from the non-empty collection \( \mathcal{C} \) of subsets of \( S \) to the real number line \( \mathbb{R} \).

In the T-s case study, the elements \( s \) of \( S \) look like \( \text{NNNTNT} \) and the rv \( I \) counts how many Ts they contain.
For instance, \( I(TNNTN) = 2 \) and \( I(NNNNTT) = (\sigma \rho / 2 \) (i.e., \( I \) ignores the order of the children). We can use the following notation to simplify things:

Notation: \[ P(\mathcal{Y} = y) = P(\{ s : I(s) = y \}) \]

For example, \( P(\mathcal{Y} = 1) = P(\{ s \in S : I(s) = 1 \}) = 2 P(\{ TNNNN, NTTNN, NNNTN, NNNNTT \}). \]

In general, the value a random variable takes on could be just about anything, but in this course all of our rvs will be real-valued.

(7 Aug 17) In the T-S case study, the rv \( I \) can only take on the values \( 0, 1, \ldots, 5 \).
You can see that a \( \mathbb{W}_2 \) is completely specified by two things: the values it can take on, and the probabilities for those values.

\[ \begin{array}{c|c}
 y & P(Y = y) \\
 0 & 0.237 \\
 1 & 0.396 \\
 2 & 0.264 \\
 3 & 0.088 \\
 4 & 0.015 \\
 5 & 0.001 \\
\end{array} \]

(see p. 13)

**Definition:** The (probability) distribution of a random variable \( Y \) is the collection of all probabilities of the form \( P(Y \in A) \) for all sets \( A \) of real numbers in the non-measurable collection \( C_{IR} \) of subsets of the real number line \( IR \).

This rv \( Y \) in the T-s case study has a finite set of possible values —
This is true of some, but not all, rvs.

**Definition:** A random variable has a discrete distribution, or equivalently "is a discrete rv," if the set of (distinct) possible values is finite or at most countably infinite; rvs for which the set of possible values is uncountable are called continuous random variable.

**Examples:**

1. The rv $X = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases}$ (with $Y = \# T-S$ tally) is discrete, taking on only the values $\{0, 1\}$. Such rvs are called dichotomous or binary.
Imagine a scale for weighing things that has a dial you can set to specify how many significant figures of precision you want. Buy a "1 pound" package of butter at your favorite market and weigh it.

If there's no conceptual limit to the number of sigfigs you could get, a rv \( X = \text{(the actual (true) weight of the package)} \) should be modeled as continuous, having values (e.g.) on \((0, \infty)\), the positive part of \( \mathbb{R} \).

Reality check: Infinite precision is impossible in practice.
every measurement you ever make is in actuality discrete, but it's useful to regard rvs that are conceptually continuous (i.e., no limit in principle to the precision of measurement) as continuous.

**Definition**

Given a discrete rv $X$, the probability function (pmf or pf) of $X$ is the function $f$ that keeps track of the probability associated with $X$: $f(y) = P(X = y)$.

The set $\{ y : f(y) > 0 \}$ is called the support of (the distribution of) $X$.

(ES is almost unique in using "pf"; nearly everybody talks about the pmf.)