Law of total probability:
\[ P(A) = \sum_{j=1}^{k} P(B_j) P(A \mid B_j) \]

These probabilities add up to 1

Expressing the probability of \( A \) as a weighted average of the conditional probabilities of \( A \) given the partition sets weighted by the probabilities of those partition sets.

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

Def: Events \( A, B \) are independent iff
\[ P(A \cap B) = P(A) \cdot P(B) \]
which (as long as \( P(A) > 0 \) and \( P(B) > 0 \)) is equivalent to the Bayesian view that
\[ P(A \mid B) = P(A) \text{ and } P(B \mid A) = P(B) \]

Consequences of the def. 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 def. to more than 2 events:
Def: Given events \( A_1, \ldots, A_k \), they are mutually independent if for every subset \( A_{i_1}, \ldots, A_{i_j} \) of \( (A_1, \ldots, A_k) \) (\( j = 2, \ldots, k \)),
\[ P(A_{i_1} \cap \ldots \cap A_{i_j}) = P(A_{i_1}) \cdots P(A_{i_j}) \]
Bayesian Interpretation of Independence: A, B are independent iff information about A doesn’t change the chances associated with B, and vice versa.

Def: Another useful extension of independence: Events \( \{A_1, ..., A_k\} \) are conditionally independent given event B if for every subset \( \{A_{i_1}, ..., A_{i_j}\} \) of \( \{A_1, ..., A_k\} \) (\( j = 2, ..., k \)),

\[
P(A_{i_1} \cap ... \cap A_{i_j} | B) = \prod_{\ell=1}^{j} P(A_{i_{\ell}} | B)
\]

Statistical Ex:

Suppose there is a machine that can take an ordinary coin and produce IID tosses of the coin with \( P(H) = \theta \) with \( H \) being any one toss, 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 \( \Sigma_1, \Sigma_2, ... \)

Suppose the first 10 tosses come at 1011100111

\( (7 H, 3 T) \) ”bits“ (binary digits)

Q: Is there information in these first 10 tosses that helps you predict the next outcome \( \Sigma_{11} \)?

A: Yes, definitely. It looks like \( \theta \) is around \( \frac{7}{10} \) so you would predict \( \Sigma_{11} = H \). Thus \( \Sigma_{11} \) depends on \( \Sigma_1 \) through \( \Sigma_{10} \) probabilistically.
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 come out HHTHTHTHHH (8H, 2T)

Q: Is there information in these 10 tosses that helps you predict the next toss?

A: No. You know that $\Theta = 0.81$ so there's no information in any of the $Y_i$ that helps you to predict any of the other $Y_j$. Given $\Theta$, the $Y_i$ are independent.

Thus, the $Y_i$ are unconditionally independent but conditionally dependent given $\Theta$.

Ex with $n$ $Y_i$ values: Graph Theory

Edges denote any dependence amongst the nodes

$Y_i$ dependent

$n$ nodes $\times n$ edges

$\frac{n(n-1)}{2}$ edges = $O(n^2)$

This is of order $n^2$ "Big O" of $n^2$

$x \rightarrow y$ causes $y$

As soon as we know what $\Theta$ is, the $Y_i$ become conditionally dependent given $\Theta$.

The past and future are conditionally independent given the truth ($\Theta$)
Bayes' Theorem for events based on the possibility of a finite partition:

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 $i = 1, \ldots, k$:

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

and by the LTP, this is:

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

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 $B_i$.

(A priori) — Before the dataset $A$ arrives, you have background information about the possibility of the $B_i$, that you can represent with prior probabilities $P(B_i)$.

(A posteriori) — After the dataset $A$ arrives, you can use Bayes' 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 among likelihood information.
The denominator \( P(A) \) does not depend on the \( B_i \), and can therefore be regarded as a normalizing constant put into Bayes' Theorem to make all the \( P(B_i \mid A) \) add up to 1.

Thus \( \frac{P(B_i) P(A \mid B_i)}{P(A)} \) is interpreted as

\[
\text{(posterior information)} = \frac{\text{(prior information)} \cdot \text{(likelihood information)}}{\text{(normalizing constant)}}
\]

Back to credit card case study:

We know \( P(B) = 0.01 \) \( P(S-1|G) = 0.97 \) \( P(S+1|B) = 0.98 \)

We want \( P(B|S+) \quad \uparrow \quad \uparrow \quad \text{true (unknown)} \quad \text{system says (data)} \quad \text{only 2 possible states: (B, G)} \)

Method 1: 2x2 table

Method 2: Bayes' Theorem in odds form

\[
\frac{P(B \mid S+)}{P(\text{not } B \mid S+)} = \left[ \frac{P(B)}{P(\text{not } B)} \right] \cdot \left[ \frac{P(S+1 \mid B)}{P(S+1 \mid \text{not } B)} \right]
\]

We know all the probabilities on the right side.

\[
\text{(posterior odds)} = \left( \frac{\text{prior odds}}{\text{Bayes factor}} \right) \quad (\text{in favor of } B)
\]

\[
\begin{align*}
P(B) &= 0.01 \\
P(\text{not } B) &= P(G) = 0.99 \\
P(S+1|B) &= 0.98 \\
P(S+1|\text{not } B) &= 1 - P(S-1|\text{not } B) = 1 - P(S-1|G) = 1 - 0.97 = 0.03
\end{align*}
\]
\[
\begin{align*}
&= \left( \frac{99}{99} \text{ to } 1 \right) \text{ (odds against B)} \left( \frac{98}{98} \text{ to } 3 \text{ odds} \right) \\
&= \frac{98}{(99)(3)} = \frac{98}{297} \quad \text{← posterior odds in favor of B} \\
&= \text{o = odds} \quad \text{o} = \frac{p}{1-p} \quad \Rightarrow \quad p = \frac{297}{98} \\
&= \frac{98}{1 + \frac{98}{297}} = \frac{98}{98 + 297} = \frac{98}{395} = 0.25 \text{ as before}
\end{align*}
\]

Method 3: Bayes' Theorem in probability form
\[
P(B|S+) = \frac{P(B)P(S+|B)}{P(S+)} = \frac{(0.01)(0.98)}{?}
\]

Truth: \{B, G\} (unknown)
Data: \{S+, S-\}

\[\text{Search for truth on basis of data: statistical inference} \]
\[\text{truth \rightarrow data (observable)} \]
\[\text{unknown (true) state of world may well be unobservable} \]

We don't know \(S^+\) by itself so we need \(B\) or \(G\)

When \(S\) is hard, get help by extending the conversation.
Find some other aspect of world \(X\) upon which \(S\) depends and predict \(X\) in 2 stages:

\[
\begin{align*}
P(S+) &= \text{data (X)} \\
\text{other info (X)} &= \{G, B\}
\end{align*}
\]

\[\text{Partition the world} \]
\[ P(S+) = P(S+ \text{ and } G) \lor P(S+ \text{ and } B) \]
\[ = P(S+ \text{ and } G) + P(S+ \text{ and } B) \]
\[ = P(G)P(S+|G) + P(B)P(S+|B) \]
\[ = (0.99)(1 - P(S-|G)) + (0.01)(0.98) \]
\[ = 0.99(1 - 0.97) + (0.01)(0.98) \]
\[ = 0.099(0.03) + (0.01)(0.98) \]
\[ = 0.0297 + 0.0098 = 0.0395 \]

\[ P(B|S+) = \frac{P(B)P(S+|B)}{P(S+)} = \frac{(0.01)(0.98)}{0.0395} = \frac{98}{395} = 0.25 \]

---

Ch. 3 Random variables and their distributions

**Example:** Tay-Sachs Disease

- \( T \) = T-S baby
- \( N \) = not T-S baby

<table>
<thead>
<tr>
<th>S</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNNNNN</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>TNTNNN</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>NNTNNT</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>NNNNTN</td>
<td>N</td>
<td>T</td>
</tr>
<tr>
<td>NNNNTT</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TTTNTT</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Def:** Given a sample space \( S \) for an experiment \( E \), a (real-valued) random variable (rv) is a function from the non-empty collection \( C \) of subsets of \( S \) to the real number line \( \mathbb{R} \).

In the T-S case study, the elements \( S \) of \( S \) look like \( NNNNTT \) and the rv \( \mathbb{Y} \) counts how many Ts they contain.

For instance, \( \mathbb{Y}(TNNNTN) = 2 \) and \( \mathbb{Y}(NNNTTT) = 2 \) also. \( \mathbb{Y} \) ignores the order of the children.

To simplify: \( P(\mathbb{Y} = \gamma) = P(\{S: \mathbb{Y}(S) = \gamma\}) \)

T/F proposition set
For ex) \( P(\bar{Y} = 1) = P(\{x \in S : \bar{X}(s) = 1\}) \)

\[ = P(\{T, T, N, N, N, N, N, N, N\} \) \]

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

In the T-S case study the RV \( \bar{Y} \) can only take on the values 0, 1, \ldots, 5

\[ P(\bar{Y} = y) \]

random variable \( \rightarrow \) a possible value of \( \bar{Y} \)
(proces) \( \rightarrow \) (outcome)

<table>
<thead>
<tr>
<th>y</th>
<th>( P(\bar{Y} = y) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.237</td>
</tr>
<tr>
<td>1</td>
<td>0.396</td>
</tr>
<tr>
<td>2</td>
<td>0.264</td>
</tr>
<tr>
<td>3</td>
<td>0.088</td>
</tr>
<tr>
<td>4</td>
<td>0.015</td>
</tr>
<tr>
<td>5</td>
<td>0.001</td>
</tr>
</tbody>
</table>

You can see that a RV \( \bar{Y} \) is completely specified by 2 things: the values it can take on and the probabilities for those values.

**Def:** The probability distribution of a random variable \( \bar{Y} \) is in the collection of all probabilities of the form \( P(\bar{Y} \in A) \) for all sets \( A \) of real #s in the non-empty collection \( C_\mathbb{R} \) of subsets of the real number line \( \mathbb{R} \).

The RV \( \bar{Y} \) in the T-S case study has a finite set of possible values— this time of some, but not all, RVs.

**Def:** A random variable \( \bar{Y} \) has a discrete distribution, or equivalently \( \bar{Y} \) is a discrete RV if the set of possible distinct values of \( \bar{Y} \) is finite or at most countably infinite.
rvs for which the set of all possible values is uncountable are called continuous random variables.

Ex: The rv $X = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases}$ (with $Y = \# \text{ of babies}$)

is discrete, taking on only the values 1 (yes, no) 2)

such rv's are called dichotomous or binary.

Ex: Imagine a scale for weighing things that has a dial you can set to specify how many significant figures (sig figs) of precision you want.

Buy a "1 pound" package of butter and weigh it.

<table>
<thead>
<tr>
<th>Possible Weights</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>15.99</td>
</tr>
<tr>
<td></td>
<td>15.993</td>
</tr>
<tr>
<td></td>
<td>15.9928</td>
</tr>
</tbody>
</table>

If there's no conceptual limit to the number of sig figs you could get, a rv $\tilde{X}$ (the actual true weight of the package) should be modeled as continuous, having positive values on the real number line $\mathbb{R} \setminus \{0\}$

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