you have an variable/feature/attribute of a system and it takes on values in some specific set. The classic example is dice throwing, with the feature being the uppermost face of the dice, taking values in \{1, 2, 3, 4, 5, 6\}

you talk of the probability of a particular feature value: \( P(X = a) \)

standard frequentist interpretation is that the systems can be observed over and over again, and that the relative frequency of \( X = a \) in all the observations tends to a stable fixed value as the number of observations tends to infinity. \( P(X = a) \) is this limit

\[
P(X = a) = \lim_{N \to \infty} \frac{\text{freq}(X = a)}{N}
\]
on this frequentest interpretation you would definitely expect the sum over different outcomes to be 1, so where \( A \) is set of possible values for feature \( X \), it is always assumed that

\[
\sum_{a \in A} P(X = a) = 1
\]

typically also interested in types or kinds of outcome: not the probability of any particular value \( X = a \). Jargon for this is event

for example, the 'event' of dice throw being even can be described as \((X = 2 \lor X = 4 \lor X = 6)\)

the relative freq. of \((2\lor 4\lor 6)\) is by definition the same as the \((\text{rel.freq } 2) + (\text{rel.freq } 4) + (\text{rel.freq } 6)\). So its not surprising that by definition the probability of an 'event' is the sum of the mutually exclusive atomic possibilities that are contained within it (ie. ways for it to happen) so

\[
P(X = 2 \lor X = 4 \lor X = 6) = P(X = 2) + P(X = 4) + P(X = 6)
\]

there's a common-sense 'explanation' for the definition

\[
P(A|B) = \frac{P(A \land B)}{P(B)}
\]

you want to take the limit as \( N \) tends to infinity of

\[
\lim_{N \to \infty} \left( \frac{\text{count}(A \land B) \text{ in } N}{\text{count}(B) \text{ in } N} \right)
\]

you get the same thing if you divide top and bottom by \( N \), so

\[
\lim_{N \to \infty} \left( \frac{\text{count}(A \land B) \text{ in } N}{\text{count}(B) \text{ in } N} \right) = \lim_{N \to \infty} \left( \frac{(\text{count}(A \land B) \text{ in } N)}{\text{count}(B) \text{ in } N} \right) / N = \lim_{N \to \infty} \left( \frac{P(A \land B)}{P(B)} \right)
\]

obviously given the definition of \( P(A|B) \), you have the obvious but as it turns out very useful

\[
P(A \land B) = P(A|B)P(B)
\]

since \( P(A|B)P(B) = P(B|A)P(A) \), you also get the famous

\[
P(A|B) = \frac{P(A \land B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}
\]
Alternative expressions of independence

- recall independence was defined to \( P(A \land B) = P(A) \times P(B) \). Given the definition of conditional probability there are equivalent formulations of independence in terms of conditional probability:
  
  **Independence**: \( P(A|B) = P(A) \)
  
  **Independence**: \( P(B|A) = P(B) \)

NOTE: each of these on its own is equivalent to \( P(A \land B) = P(A) \times P(B) \)

- Now suppose you have more than one feature/attribute of the system/situation eg. you roll two dices, use \( X \) for first, \( Y \) for second
  - can specify events with values on both variables eg. \{ \( X = 1 \), \( Y = 2 \) \} and the probability of such an event is called a joint probability
  - if \( A \) is range of values for \( X \), \( B \) range for \( Y \) must have
    \[
    \sum_{a \in A, b \in B} P(X = a, Y = b) = 1
    \]
  - can specify events with value on just one feature
    \{ \( X = 1 \) \} or maybe \{ \( Y = 2 \) \}
  - the probabilities of these are called marginal probabilities and are obtained by summing the joints with all possible values of the other feature
    \[
    P(X = 1) = \sum_{b \in B} P(X = 1, Y = b)
    \]
  
  - you say \( P(X|Y) = P(X) \) and the features \( X \) and \( Y \) are independent in case for every value \( a \) for \( X \) and \( b \) for \( Y \) you have
    \[
    P(X = a, Y = b)/P(Y = b) = P(X = a)
    \]

Chain Rule

- generalising to more variables, you can derive the indispensable chain rule

\[
P(X, Y, Z) = p(Z|(X, Y)) \times P(X, Y) = p(Z|(X, Y)) \times P(Y|X) \times p(X)
\]

\[
P(X_1 \ldots X_n) = p(X_n|(X_1 \ldots X_{n-1})) \times \ldots \times p(X_2|X_1) \times p(X_1)
\]

important to note that this chain-rule re-expression of a joint probability as a product does not make any independence assumptions
Conditional Independence

- there is a notion of **conditional independence**. It may be that two variables $X$ and $Y$ are not in general independent, but given a value for a third variable $Z$, $X$ and $Y$ become independent.

### Conditional Indpt

\[ P(X, Y|Z) = P(X|Z)P(Y|Z) \]

- as with straightforward independence there is an alternative expression for this, stating how a conditioning factor can be dropped

### Conditional Indpt altern. def

\[ P(X|Y, Z) = P(X|Z) \]

- Real-life cases of this arise where $Z$ describes a *cause*, which manifests itself into two *effects* $X$ and $Y$, which though very dependent on $Z$, do not directly influence each other
- The theories behind Speech Recognition and Machine Translation typically make a lot of *conditional independence* assumptions