Probability in Plain Language

PHS Launch!

Anna Siefkas (alynnpalevsky@g.harvard.edu)

August 25, 2022

Outline

Big picture

First, a short video...

Key points

- Intuition about probability
- Probability distributions
- Independent vs dependent events

Why is probability useful?

Probability allows us to make concrete statements about whether or not an event is likely to occur, while accounting for the fact that randomness exists.

When applied in a formal context, probability is a key tool that allows us to make **inference** [1] both about relationships between variables and about whether those relationships are due to random variation or something more causal.

[1] see the rest of PHS2000A for more!

Definitions

Probability

Intuitive definition?

Probability

Technical definition: a quantitative (i.e., numerical) measure of how likely a given event is to occur. Probability is quantified as a ratio of occurrences of the event of interest over all possible events, over an infinite number of attempted events. [2] Probabilities can take values between 0 and 1, inclusive.

We sometimes call the "occurrence of the event of interest" a "success."

[2] Adapted from Rosner. (1995). Fundamentals of biostatistics (4th ed.). Duxbury Press.

Question:

Why would we quantify probability over an infinite number of attempted events?

Random variables

A **random variable** can be thought of as any event or measurement (e.g, rolling a dice, taking a blood pressure reading) that has probability associated with it. That is, the exact value the random variable will take is not known with certainty before the event or measurement.

We can develop more technical definitions of random variables as those that have a probability mass or density function associated with them. In other words, the possible values of a random variable follow a (possibly knowable) distribution of values, with some values possibly more common than others.

Random variables

Random variables can be either **discrete** or **continuous**.

- Discrete: the variable can only take on integer (1, 2, 3, ...) values, either up to some finite limit or for an infinite number of such values.
- Continuous: the variable can take on any real value (e.g., $\sqrt{2}$, 3.2, $\frac{15}{4}$, but not $\sqrt{-2}$). These values could be over some finite range (e.g., between -1 and 1) or could include all real numbers (i.e., the real line $\mathbb{R}$ from $-\infty$ to ∞).

Random variables: examples

- Can we conceive of COVID-19 case counts as a random variable? What about COVID-19 incidence rates?
- What types of random variables would these be?

Sample space

The *sample space* ($\mathcal{S}$) is the set containing all possible outcomes of a trial.

- *Question*: What does the sample space for flipping a coin contain? What about rolling a dice?

Events of interest (e.g., rolling a 6) can be thought of as a **subset** of this sample space. We denote subsets as $A \subset \mathcal{S}$, where A is one possible outcome of the sample space $\mathcal{S}$.

This framework can be helpful if you want to think visually about probabilities.

Conditional probability (the prelude)

Conditional vs Unconditional

So far, we've been implicitly talking about **unconditional** probabilities. A probability can also be conditional. A probability is conditional when we are concerned with the probability of an event **among** some group with a particular characteristic, or **among** a group who experiences a second event of interest.

Conditional probabilities are common in the quantitative methods learned in this course. Sometimes conditional probabilities are necessary to make the methods "work" and remove bias; sometimes conditional probabilities are of scientific interest in their own right.

Conditional probabilities: examples

Unconditional

- probability of being hospitalized with COVID-19
- probability of death after a heart attack
- probability of flipping a coin and landing on heads

Conditional

- probability of being hospitalized with COVID-19 **among** those who are vaccinated
- probability of death after a heart attack **among** those over age 70
- probability of flipping a coin and landing on heads **for those in** the universe in which Rosencrantx and Guildenstern live

Conditional probability notation

We notate conditional probability with a vertical bar. If we want to know the probability of A conditional on event B, we write $Pr(A|B)$.

Mathematically, we define a conditional probability as

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}$$

So what's that upside down U symbol up there?

Operating with probability

Ands and *Ors*

When we conceive of probabilities as taking place in a sample space, it's easy to use some basic set operators to think about how we can combine the probabilities of multiple events.

AND

When we are interested in the probability of the co-occurrence of two (or more!) events of interest, we look at their **intersection**.

The intersection of A and B is the subset of the sample space where both A **and** B occur.

We denote intersections with the symbol $\cap$.

OR

When we are interested in whether at least one of a subset of events of interest occur, we look at their **union**.

The union of A and B is the subset of the sample space where A **or** B **or** both occur.

We denote unions with the symbol $\cup$.

Ands and *Ors*: Venn diagrams

AND

OR

Ands and *Ors*: examples

Can you provide some public-health related examples of situations where the probabilities of an intersection of two events may be of interest?

What about situations where the union of two events is of interest?

Mathematical definitions

Intersections

$$P(A) \cap P(B) = P(A|B) \times P(B)$$
$$= P(B|A) \times P(A)$$

Why? We need to consider first the probability of the first happening, **and then** the probability of the second even happening. This "and then" happens mathematically with multiplication. There are two ways we can define "first event" and "second" event, leading to two equivalent formulas.

Unions

$$P(A) \cup P(B) = P(A) + P(B) - P(A \cap B)$$

Why? We want to count everything that is in either the A or B portion of the sample space. The portion of the sample space that is $(A \cap B)$ is part of the subset A and part of the subset B. We need to subtract this value to avoid it being counted twice in our probability.

Mathematical definitions

For any two random variables, we can define their intersections and unions as shown on the previous slide. Drawing a Venn diagram might help clarify why these hold!

There are also some special cases, or specific manipulate we can define intersections and unions that depend on the precise way that all of the events in our subset of interest relate to each other.

Two important types of events are **mutually exclusive** and **independent events**.

Mutually exclusive events

Events in a set are said to be mutually exclusive if the occurrence of one event means that no other event in the set can occur.

If event A is flipping a coin and landing on heads, while B is landing on tails, these events are mutually exclusive. This is a simple example, but helpful for thinking about how mutually exclusive events operate.

What are some less trivial (and more public health-related) examples?

Mutually exclusive events

The rules for intersections and unions of mutually exclusive events can be worked out by thinking about the coin flipping example.

Let A be the event of landing on heads, and B be the event of landing on tails. Don't worry about the exact values, but assume $P(A)$ and $P(B)$ can be defined (we're not Rosencrantz and Guildenstern here).

Intersections

$$P(A) \cap P(B) = ?$$

Unions

$$P(A) \cup P(B) = ?$$

Mutually exclusive events

Intersections

$$P(A) \cap P(B) = 0$$

Why? Since A occurring means B cannot occur, and vice versa, the probability of A **and** B occurring is 0.

This will be true for any pair of mutually exclusive events, even if they don't make up the whole sample space.

Unions

$$P(A) \cup P(B) = P(A) + P(B)$$

Why? We are interested in the probability of A occurring **or** B occurring, and know that only one can occur. We can list each event of interest and add up their probabilities.

In this case, since A and B are the only events in our sample space, we also know that $P(A) \cup P(B) = 1$.

Independent events

Two events A and B are independent (denoted by $A \perp\!\!\!\perp B$) if the probability of A occurring does not impact the probability of B occurring.

Two coin flips are independent (for us, although maybe not for Rosencrantz and Guildenstern), since the outcome of flip number 1 doesn't impact what we expect to see for flip number 2.

Question: are mutually exclusive events independent?

Independent events

Using the example of two consecutive coin flips (in our universe), we can illustrate the rules for unions and intersections of independent events.

Intersections

$$P(A) \cap P(B) = P(A) \times P(B)$$

Why? Well, this uses a fact about conditional probabilities that we'll see in just a moment.

Unions

$$P(A) \cup P(B) = P(A) + P(B) \times [1 - Pr(A)]$$

Why? Here, there's no "overlap" between A and B that we need to worry about. Instead, we can conceptualize our events as "A occurs" and "B occurs *and A does not*". $P(B) \times [1 - Pr(A)]$ captures that second event.

Independent events: a side note

The rule $P(A) \times P(B) = Pr(A \cap B)$ is sometimes used to define independent events. However, whether you use this to try and identify independent events or claim that two events are independent so that you can use this rule can get a little circuitous.

Independence is a property we need to assume for a lot of the models and methods we use in quantitative research to hold, but it's not something we can always check with this rule!

Conditional probability (for real)

Question

Is there a difference between and intersection and a conditional probability?

Are $P(A \cap B)$ and $P(A|B)$ the same expression?

Conditional probabilities vs intersections

$P(A \cap B)$ and $P(A|B)$ are *not* the same expression!

$\mathbf{P(A \cap B)}$: Venn diagram

$\mathbf{P(A|B)}$: Venn diagram

A coin flipping example helps illustrate this. Let A be heads and B be tails, and let the subscripts denote the sequence of flips we're performing.

What values do you expect for $P(A_2 \cap A_1)$ and $P(A_2|A_1)$?

Conditional probabilities vs intersections

$P(A_2 \cap A_1)$ is the probability of flipping heads on flip one **and then** flipping heads on flip two. In our universe, this has a value of $P(A_2 \cap A_1) = 0.25$.

In contrast, $P(A_2|A_1)$ is the probability of flipping heads on flip two, **among** the set of sequences that started with a heads on flip one. Since we know successive flips are independent, in our universe, this has a value of $P(A_2|A_1) = 0.5$.

Notice that in the example above, $P(A_2|A_1) = 0.5$, where A_1 and A_2 are independent ($A_1 \perp\!\!\!\perp A_2$). In fact, this will always be true of independent random variables.

If A and B are two random variables such that $A \perp\!\!\!\perp B$, then $P(A = a) = P(A = a|B = b)$ for all possible values of b. This is a useful identity that you will use a *lot* in this class and others!

Flipping the conditioning statement

In general, it is unfortunately true that $P(A|B) \neq P(B|A)$. Why is that an unfortunate fact?

It's fairly common to only have information on one of those quantities, when the other quantity is of greater public health or scientific interest.

What are some common examples of this conundrum?

Bayes' Theorem

Bayes' Theorem allows us to go from $P(A|B)$ to $P(B|A)$. We start with the fact that

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ which we can rewrite as } P(A \cap B) = P(A|B)P(B).$$

We also know that

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

We can plug the right hand side of the expression above into this to get the following formula for $P(B|A)$:

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

The more general form of this theorem is shown in the supplemental slides.

Probability functions

Probability functions

Several functions exists which help us to summarize information about the probability of all possible outcomes in the sample space, all at once.

These functions can provide graphical information on the probability of various events, as well as provide information on other important aspects of the probability distribution (such as its variance).

We'll use two classes of such functions regularly: 1) the probability mass function or probability density function and 2) the cumulative distribution function.

Probability mass functions

A probability mass function (PMF) is used to describe the probability associated with each possible value that a discrete random variable can take. We notate the PMF as $P(X = k)$ or more generally as $f_X(x)$.

For this to work, we generally need to define our discrete random variable in terms of a set of numerical values. Sometimes these numerical values are obvious (e.g., rolling a dice) and sometimes they are arbitrary (e.g., flipping a coin). We notate the numerical values for a random variable as:

$$X = \begin{cases} 1 & \text{if we get heads} \\ 2 & \text{if we get tails} \end{cases}$$

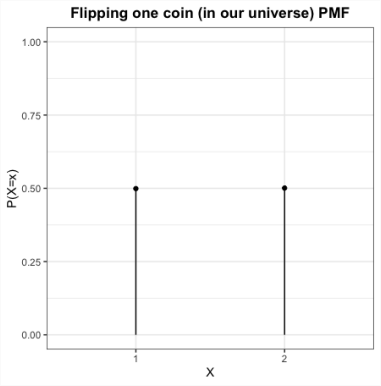
Assuming we're in our universe, working with normal coin flipping rules, what would the PMF associated with this random variable be?

Probability mass functions

Assuming we're in our universe, working with normal coin flipping rules, what would the PMF associated with this random variable be?

If we expect 50% of flips to land on heads and 50% on tails, we have:

$$P(X = k) = \begin{cases} 0.5 & x = 1 \\ 0.5 & x = 2 \end{cases}$$



Probability mass function

Can you provide another example of a discrete random variable and its PMF?

Random variable? PMF values? PMF graph?

Probability mass functions

Some discrete random variables follow known probability distributions, and have more complicated functions defining their PMFs. Some common examples are in the supplemental slides.

No matter what the random variable is or how many discrete values it can take on, the cumulative height of all the function at all defined points must add up to one.

Question: Why is that?

Probability density function

We said above that the PMF was used to describe discrete random variables.

Question: Why can't we use the PMF for continuous random variables as well?

Probability density function

In a sense, the probability *density* function (PDF) is what happens when you add points at every single value along the x-axis. In the case of the PDF, the total area under the curve will always be equal to one. As with the PMF, the PDF is notated as $f_X(x)$.

However, with a continuous random variable, the probability of the variable taking on any specific value is 0.

- For this reason, we generally define PDFs in terms of the density (area under the curve) less than, greater than, or between some values(s) of interest. Occassionally, these can be read off the PDF itself.

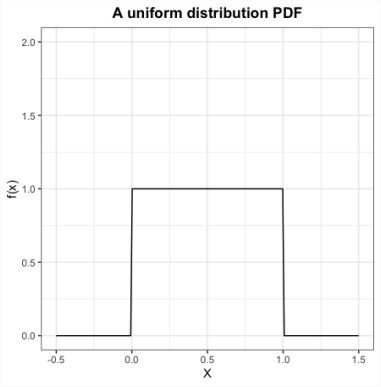
Probability density function

For the (very, very exciting) PDF on this slide, consider:

- Why can the Y-axis value be 1 for more than one value of X?
- What could this distribution represent in the real world?

And for the math:

- What is the total area under this PDF?
- What is $P(X \leq 0.8)$?
- What is $P(0.4 \leq X \leq 0.6)$?



Probability density function

However, the PDF is often more complicated than the uniform function above. The PDFs for common continuous random variables are presented in the supplemental slides.

You can see for these functions that it's a little more difficult to read quantities like the ones above directly off the graph!

Cumulative distribution function

The **cumulative distribution function** (CDF), notated as $F_X(x)$, directly defines the quantity $P(X \leq x)$.

Based on the previous slides, how would you imagine the CDF is defined?

Cumulative distribution function

$$F_X(x) = \int_{-\infty}^{x} f_X(x)$$

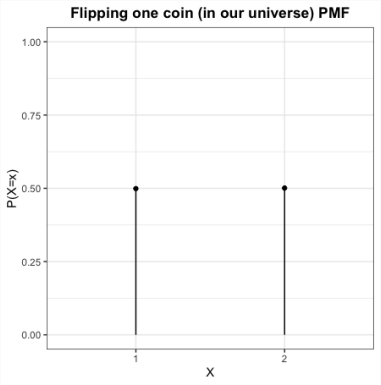
For a discrete random variable, the values of the CDF could be identified simply by summing over the values of the PMF. Similarly, for a continuous random variable, the values of the CDF could be identified by integrating over the values of the PDF.

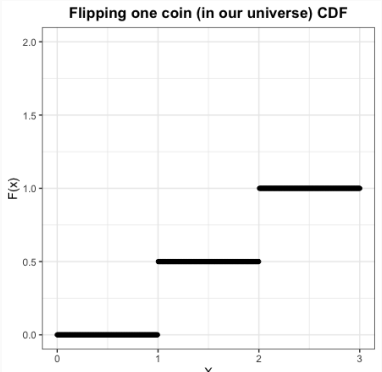
In either case, we wind up with a function $F_X(x)$ where the actual values of the function directly correspond to $P(X \leq x)$.

Cumulative distribution function: discrete example

$$f_X(x) = \begin{cases} 0.5 & x = 1 \\ 0.5 & x = 2 \end{cases}$$

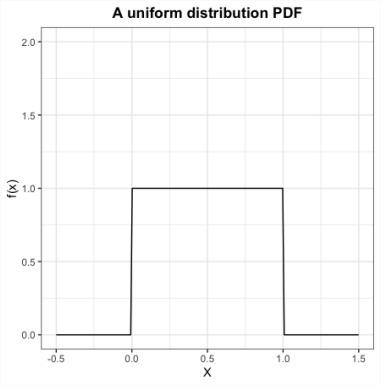
$$F_X(x) = \begin{cases} 0 & x < 1 \\ 0.5 & 1 \leq x < 2 \\ 1 & x \leq 2 \end{cases}$$



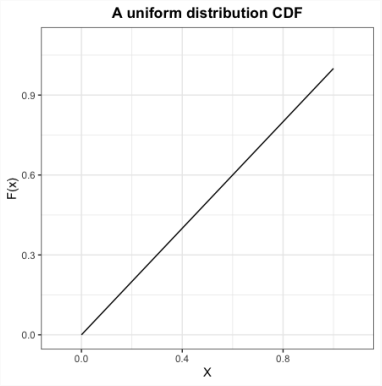


Cumulative distribution function: continuous example

$f_X(x) = 1$ for $x \in (0, 1)$



$F_X(x) = x$ for $x \in (0, 1)$



Dealing with more than one random variable

Frequently, we will have to consider the distribution of more than one variable at a time. In these cases, we can define a **joint** probability distribution (either as a PMF or PDF) as $f_{XY}(x, y)$.

In the special case where X and Y are independent random variables, this function decomposes to

$$f_{XY}(x, y) = f_X(x) \times f_Y(y).$$

This may be a useful property to keep in mind tomorrow and onwards!

That's it!

Thank you!

Many thanks to Yiwen Zhu (TF for PHS2000A 2021) for creating the previous version of this presentation, and to the rest of the PHS2000A teaching fellows for comments, edits, and sense-checks on these slides.

Supplemental slides

Key distributions

Bernoulli and binomial distributions

Both describe trials of events with probability p of success.

Bernoulli distribution

- Describes a single trial of an event with probability p.
- The probability distribution defines the probability that the single trial is a "success".

Binomial distribution

- Describes n trials of an event with probability p, where n is an integer.
- The probability distribution defines the probability that k out of the n trials are a "success".

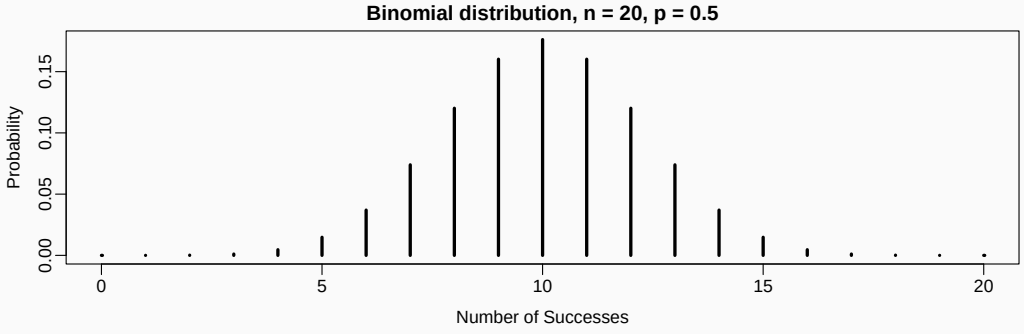
Bernoulli and binomial distributions

Bernoulli distribution

- Parameterized with p alone.
- The PMF takes the form $P(X = k) = p^k(1-p)^{1-k}$ where $k \in \{0, 1\}$

Binomial distribution

- Parameterized with p and n.
- The PMF takes the form $P(X = k) = \binom{n}{k} p^k(1-p)^{n-k}$ where $k \in \{0, 1, 2, ..., n\}$.



Poisson distribution

The Poisson distribution is used to describe the distribution of *count* events. Examples may include - Traffic accidents at a particular intersection - CVD events in a population - Times my cat knocks a pen off the table

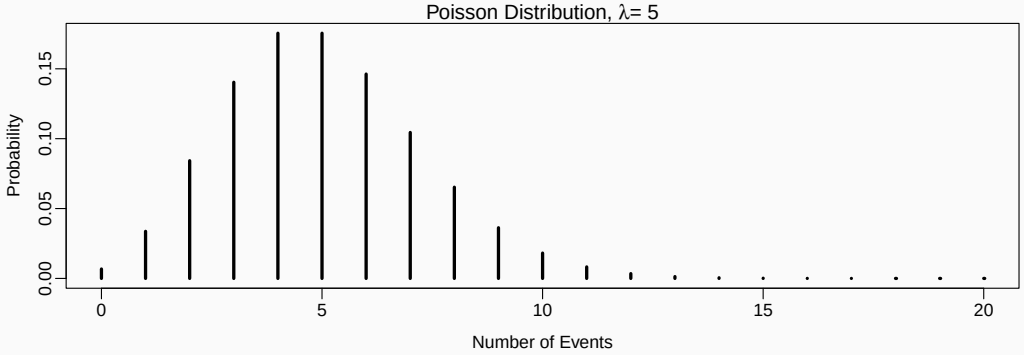
Processes that give rise to events that follow a Poisson distribution are often called *Poisson processes*.

When we consider counts generated by a Poisson process occurring *over time*, Poisson distributions can help us to understand and model event **rates**.

Poisson distribution

The Poisson distribution can be fully described by the parameter λ, which is both the expected count (mean) and the variance of the distribution.

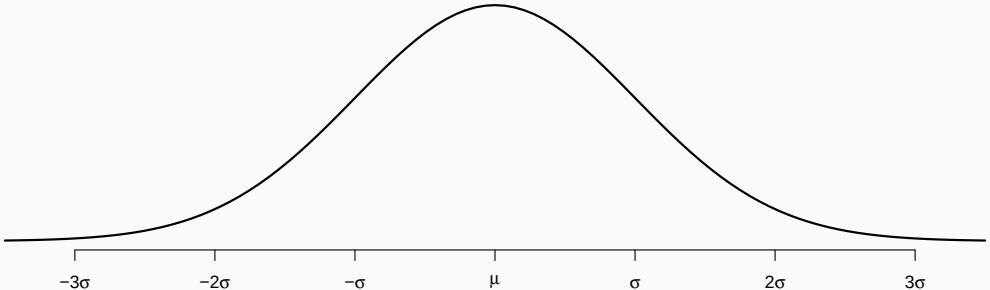
The PMF takes the form $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ where $k \in \{0, 1, 2, ..., n\}$



Normal distribution

Describes continuous data with values spread evenly to both sides of the mean

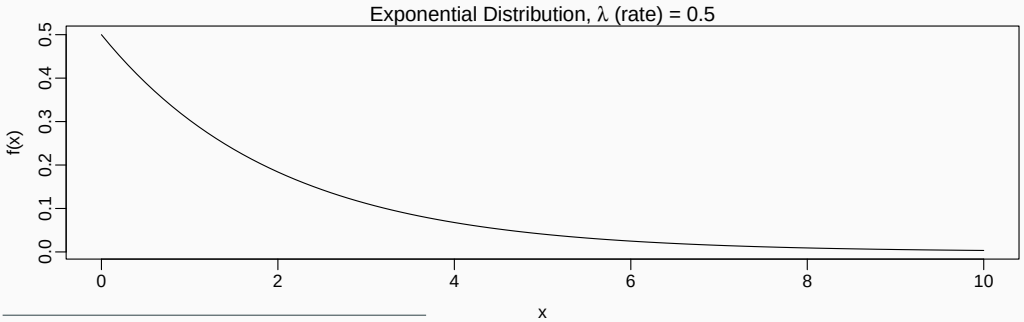
- Can be complete defined with mean μ and variance σ^2.
- Possible values range from $-\infty$ to ∞ (although there is very little density in the tails!)
- The PDF takes the form $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$



Exponential distribution

The exponential distribution describes the time between events, where the events are generated by a Poisson process. The distribution can be parameterized by a single parameter λ, which here denotes the rate (hazard) at which events occur.[3]

The PDF takes the form $f_X(x) = \lambda e^{-\lambda x}$ where $x \in (0, \infty)$



[3] See the survival lectures later this fall for more on the hazard!

Useful probability rules

Bayes' Theorem

If $A_1, A_2, ...A_n$ are mutually exclusive and exhaustive subsets of a sample space (meaning at least one of $A_1, ..., A_n$) must always occur), then for each value $i \in \{1, ..., n\}$,

$$P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{\sum_{j=1}^{n} P(B|A_j) \times P(A_j)}$$

Note that we call $A_1, A_2, ...A_n$ a *partition of the sample space*, and notate it as $\bigcup_{i=1}^{n} A_i = S$.

Law of total probability

The law of total probability (or total expectation) allows us to "condition out" or "marginalize over" a variable that we do not want to condition on.

$$P(A) = \sum_x P(A|X = x) \times P(X = x)$$

This also holds for conditional probabilities.

$$P(A|B) = \sum_x P(A|B, X = x) \times P(X = x|B)$$

In the notation above, $\sum_x$ is commonly used shorthand for "sum over all possible values of the random variable $X = x$. If X is a standard six-sided dice, $\sum_x$ would mean summing over the values $X = 1, X = 2, ..., X = 6$, for example.

Rule of iterated expectations

You'll probably use this one less than the law of total probability above, but it may be useful to know. It holds both for probabilities and for expectations more generally.

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