Bayesian Modeling

Interpretation of probabilities and Bayes' formulas

Yixin Wang

Preliminary Draft.
Please do not distribute.

2 Interpretation of probabilities and Bayes' formulas

2.1 Interpretation of probabilities

When we say: "if I toss a coin, the probability that the coin turns head is 1/2", what we understand is the possibility of repeated experiments of coin tossing, and approximately half of the times the coin turns head. This is the frequentist interpretation of probabilities. This is also the interpretation we rely on when we think of sampling model $\Pr(Y = 1|\theta = 1/2)$.

But what do we mean by saying, a day before the votes are cast, that a candidate wins the election with probability 66.67%? We cannot repeat the election multiple times for the same candidates. This probability number obviously quantifies the degree of our belief in a given statement. The higher the number, say 80, or 95%, the stronger the belief (which is subjective, since my 80% may be differently perceived from your 80%). We use this interpretation when we specify the prior $\Pr(\theta)$ and drawing inference from the posterior distribution $\Pr(\theta|\text{Data})$.

Both interpretations are present in Bayesian analysis in the prior and the sampling model terms and linked via the Bayes formula. More remarkably, the Bayes formula enables us to revert the arguments in *conditional probabilities*, i.e., to relate $\Pr(A|B)$ with $\Pr(B|A)$, and so on. We can makes sense of and quantifies both statements such as "if a person has college degree, then his likely income level is...", versus "if a person has this income level, then they are likely to have received a college degree".

In logic, it is simple to distinguish the logical statements $A \Rightarrow B$ and $B \Rightarrow A$. In probabilistic settings and real-life applications, it is not so obvious to quantify the uncertainty of such statements.

2.2 Bayes' formulas

Bayes' formulas are straightforward to grasp in the somewhat abstract language of probability space and Venn diagrams of subsets of events. Later, we apply Bayes' rule to random variables, as commonly done in practice. (Paradoxically, the application of Bayes' rule to random variables seems less intuitive in specific applied settings).

Let $\mathcal{H}$ be the set of all possible truths, that we can place the unit probability on: $\Pr(\mathcal{H}) = 1$. Suppose $\{H_1, \ldots, H_K\}$ be a partition of $\mathcal{H}$. The rule of total probability imposes that

$$\sum_{k=1}^{K} \Pr(H_k) = 1.$$

Examples

- $\mathcal{H}$ is the set of truths about people's religious orientations. Partitions include {Christian, non-Christian}, but also {Protestian, Catholic, Jewish, other, none}, and so on.
- $\mathcal{H}$ is the set of truths about people's number of children.
- $\mathcal{H}$ is the set of truths about the relationship between smoking and hypertension in a given population. Partitions include {some relationship, no relationship}, or
  {negative correlation, zero correlation, positive correlation, and so on.

An even E is defined as a subset of $\mathcal{H}$ for which we may quantify in terms of $\Pr(E)$. By the rule of marginal probability:

$$\Pr(E) = \sum_{k=1}^{K} \Pr(E \cap H_k) = \sum_{k=1}^{K} \Pr(H_k) \Pr(E|H_k),$$

where we have used the definition of conditional probability in the second equality.

It follows that

$$\begin{aligned} \Pr(H_j|E) &= \frac{\Pr(H_j \cap E)}{\Pr(E)} \\ &= \frac{\Pr(E|H_j) \Pr(H_j)}{\sum_{k=1}^{K} \Pr(E|H_k) \Pr(H_k)}. \end{aligned}$$

This is an instance of the celebrated Bayes' formulas, which allows one to compute the "inverse probability" $\Pr(H_j|E)$ in terms of $\Pr(E|H_j)$ and other quantities. The *other* quantities here are the seemingly benign *unconditional probability* terms $\Pr(H_j)$. In reality it is often the presence of understated or hidden assumptions about these conditional probabilities that lead people to draw drastically contradictory conclusions in the face of the same set of observed evidence. Bayes' formulas explain this phenomenon clearly.

Example 2.1. A subset of the 1996 General Social Survey includes data on the education level and income for a sample of males over 30 years of age. Let $\{H_1, H_2, H_3, H_4\}$ be the events that a random selected person in this sample is in the lowest, the second, the third and the upper 25th percentile in terms of the income. By definition, the unconditional probabilities are

$$\{\Pr(H_1), \Pr(H_2), \Pr(H_3), \Pr(H_4)\} = \{.25, .25, .25, .25\}.$$

These probabilities add up to 1.

Let E be the event that a randomly sampled person from the survey has a college education. From the survey data, we also have

$$\{\Pr(E|H_1), \Pr(E|H_2), \Pr(E|H_3), \Pr(E|H_4)\} = \{.11, .19, .31, .54\}.$$

These are also probabilities. They do not add up to one. Rather, they represent the proportions of college degree holders in each of the four subpopulations. Observe the increase in the proportion relative to the income percentile level.

Now, applying Bayes' rule to obtain

$$\{\Pr(H_1|E), \Pr(H_2|E), \Pr(H_3|E), \Pr(H_4|E)\} = \{.09, 0.17, .27, .47\}.$$

What we see here are the probability that someone is in each of the income basket, if that person is a college degree holder. These probabilities add up to one. Note how the share the same monotonicity with the numbers in the previous paragraph. This is by design, because the unconditional probabilities $\Pr(H_i)$ are the same. The monotonicity will not be preserved in general and may be counterintuitive, if the subpopulations $\{H_i\}$ are partitioned such a way that their corresponding probabilities

$$\{\Pr(H_1), \Pr(H_2), \Pr(H_3), \Pr(H_4)\}$$

are suitably skewed [Exercise: come up with an example!]

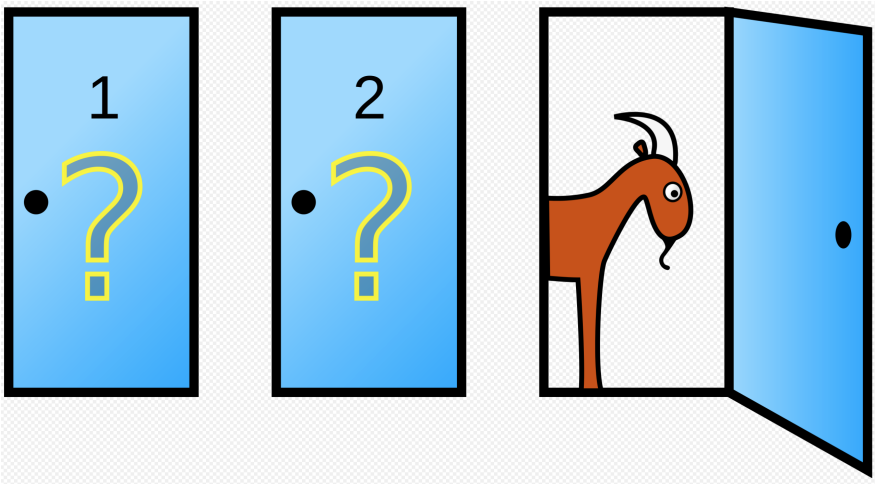

Figure 2.1: Monty Hall problem (Source: Wikipedia).

Example 2.2. (Game of chances) You are on a game show, and given the choice of three doors. Behind one door is a car, behind the others, goats. Suppose that you pick a door, say No. 1 and the host, who knows what's behind the doors, then opens another door, say No. 3, which has the goat.

He then says to you: "Do you want to switch the door, or stick to your original one?"

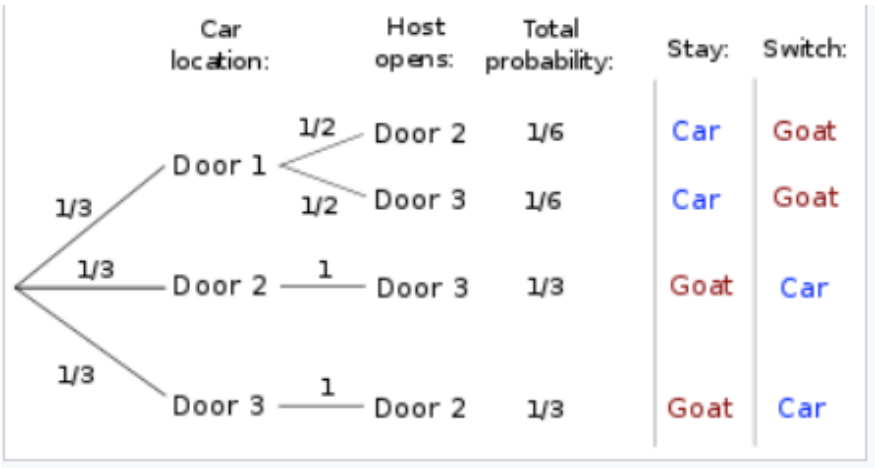

Figure 2.2: Tree representing all possibilities.

Without acting on the additional information, we know

$$\Pr(\text{Car behind Door 2}) = \Pr(\text{Car behind Door 1}) = 1/3.$$

The probability of winning by switching the door:

$$\begin{aligned}
\Pr(\text{Car behind Door 2} \mid \text{Host opens Door 3}) &= \frac{\Pr(\text{Car behind Door 2} \cap \text{Host opens Door 3})}{\Pr(\text{Host opens Door 3})} \\
&= \frac{(1/3) \times 1}{(1/3) \times 1 + (1/3) \times (1/2)} \\
&= \frac{1/3}{1/2} = 2/3.
\end{aligned}$$

This implies that $\Pr(\text{Car behind Door 1} \mid \text{Host opens Door 3}) = 1 - 2/3 = 1/3$. So,

$$\Pr(\text{Car behind Door 2} \mid \text{Host opens Door 3}) > \Pr(\text{Car behind Door 1} \mid \text{Host opens Door 3}).$$

Although the car's position has not changed, our belief about its position has, given the new information!

2.3 Bayesian hypothesis testing

In Bayesian inference, $\{H_1, \ldots, H_K\}$ often refer to disjoint hypotheses or states of nature, and E refers to the outcome of the survey, study or experiment. To compare the hypotheses post-experimentally, we may calculate the ratio

$$\begin{aligned} \frac{\Pr(H_i|E)}{\Pr(H_j|E)} &= \frac{\Pr(E|H_i)}{\Pr(E|H_j} \times \frac{\Pr(H_i)}{\Pr(H_j)} \\ &= \text{"Bayes factor"} \times \text{"prior beliefs"}. \end{aligned}$$

This tells us that the Bayes' rule only tells us what our beliefs should be after seeing the data; the prior beliefs play a very important role. The following example is apt given the most recent election:

$$\begin{aligned} \mathcal{H} &= \text{all possible rates of support for candidate A} \\ H_1 &= \text{more than half the voters support candidate A} \\ H_2 &= \text{less than or equal to half the voters support candidate A} \\ E &= \text{54 out of 100 people surveyed said they support candidate A} \end{aligned}$$

In the face of the polling data E, how should we conclude about the chance of candidate A? The modeling of both $\{\Pr(H_i)\}$ and $\Pr(E|H_i)$, and the interplay among these quantities combine to determine the inference.

2.4 Random variables and conditional independence

Bayesian inference is applied to random variables: the observed data y and the quantity of interest θ are both realizations of random variables.

The domain of these random variables and the structural properties about them have to be taken into account in order to construct suitable probability models for which the Bayes formula can be applied.

2.4.1 Discrete domains

We say Y is discrete if its domain $\mathcal{Y}$ is countable, meaning that it can be expressed as $\mathcal{Y} = \{y_1, y_2, \ldots\}$.

The event that the outcome Y takes a value y can be quantified by the probability $\Pr(\{Y = y\}) := p(y)$, where function p is called *probability mass funciton* (pmf), or *probability density function* (pdf) of Y. It satisfies the property that

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$.
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$.

An event of interest concerning the outcome Y takes the form $Y \in A$, for some subset $A \subset \mathcal{Y}$. We may quantify our belief about such an event via

$$\Pr(Y \in A) = \sum_{y \in A} p(y).$$

There are many examples of probability distributions on discrete domains. They will form crucial building blocks we will need for the probability models we will construct. Here are a few examples; it is important to review them.

1. bernoulli$(y|\theta)$, where $y \in \{0, 1\}, \theta \in [0, 1]$. The pdf takes the form
$$p(y|\theta) = \theta^y (1-\theta)^{1-y}.$$

2. binomial$(y|\theta, n)$, where $y \in \mathbb{N}$ and $\theta \in [0, 1]$.
$$p(y|\theta) = \binom{n}{y} \theta^y (n-\theta)^{1-y}.$$

3. poisson$(y|\theta)$, where $y \in \mathbb{N}$, $\theta \geq 0$.
$$p(y|\theta) = \theta^y e^{-\theta}/y!.$$

4. categorical$(y|\theta)$, where $y \in \{1, \ldots, K\}$, $\theta \in \Delta^{K-1} := \{(q_1, \ldots, q_K) \in \mathbb{R}_+^K, \sum_{k=1}^K q_k = 1\}$.
$$p(y|\theta) = \theta_y = \prod_{k=1}^K \theta_k^{\mathbb{I}(y=k)}.$$

5. multinomial$(y|\theta, n)$, where $y = (y_1, \ldots, y_K) \in \mathbb{N}^K$ such that $\sum_{k=1}^K y_k = n$, $\theta \in \Delta^{K-1}$.
$$p(y|\theta, n) = \binom{n}{y_1 \ldots y_K} \prod_{k=1}^K \theta_k^{y_k}.$$

We have used Y to illustrated random variables of discrete domains, but remember that in Bayesian inference, the quantity of interest θ is also random, for which we apply the prior distributions that are drawn from the same tool box as mentioned.

2.4.2 Continuous domains

By this, we mean the domain of the variable is the real line or a subset of the real line. We have a rich tool box of modeling devices, including distributions by the name of Gauss, Laplace, Cauchy, Gamma, Beta, Dirichlet, and so on, and beyond. Many of these building blocks can be viewed as instance of distributions in the exponential families of distributions. We will return to this in the sequel.

2.4.3 Multivariate domains

Most interesting and challenging scenarios deal with multiple variables and/or variables of multiple dimensions. How do we specify probability distributions in these cases?

Let us start with bivariate distributions in a discrete domain. Consider discrete random variables Y_1 and Y_2 taking values in countable spaces $\mathcal{Y}_1, \mathcal{Y}_2$, respectively. We need to specify the joint probability density function (joint pdf):

$$p_{Y_1Y_2}(y_1, y_2) := \Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}). \tag{2}$$

If Y_1 and Y_2 are mutually *independent*, the joint pdf is simplified to the product form

$$p_{Y_1Y_2}(y_1, y_2) = p_{Y_1}(y_1)p_{Y_2}(y_2),$$

where the two univariate pdf for Y_1 and Y_2 may be specified using to basic building block mentioned earlier.

In general, Y_1 and Y_2 are not independent; one needs to specify the the joint pdf in Eq. (2), which defines the probability mass for each of the $|\mathcal{Y}_1| \times |\mathcal{Y}_2|$ pairs of numerical values of (y_1, y_2).

Once the joint pdf is specified, the marginal distribution and conditional distribution can be computed from the joint density:

$$p_{Y_1}(y_1) := \sum_{y_2 \in \mathcal{Y}_2} p_{Y_1 Y_2}(y_1, y_2),$$
$$p_{Y_2|Y_1}(y_2|y_1) = \frac{\Pr(Y_1 = y_1, Y_2 = y_2)}{\Pr(Y_1 = y_1)} = \frac{p_{Y_1 Y_2}(y_1, y_2)}{p_{Y_1}(y_1)}.$$

From the above, we can alternatively specify the joint $p_{Y_1 Y_2}$ by first specifying marginal distribution, say p_{Y_1}, and then the conditional pdf $p_{Y_2|Y_1}$, because

$$p_{Y_1 Y_2}(y_1, y_2) = p_{Y_1}(y_1) p_{Y_2}(y_2|y_1) = p_{Y_2} y_2 p_{Y_1}(y_1|y_2).$$

When the context of the random variables is clear, we may drop the subscripts to write the above as

$$p(y_1, y_2) = p(y_1) p(y_2|y_1) = p(y_2) p(y_1|y_2).$$

Example 2.3. Let's start with the following example from PH (pg. 24) and then expand on this.

Example: Social mobility

Logan (1983) reports the following joint distribution of occupational categories of fathers and sons:

	son's occupation				
father's occupation	farm	operatives	craftsmen	sales	professional
farm	0.018	0.035	0.031	0.008	0.018
operatives	0.002	0.112	0.064	0.032	0.069
craftsmen	0.001	0.066	0.094	0.032	0.084
sales	0.001	0.018	0.019	0.010	0.051
professional	0.001	0.029	0.032	0.043	0.130

Suppose we are to sample a father-son pair from this population. Let Y_1 be the father's occupation and Y_2 the son's occupation. Then

$$\begin{aligned}\Pr(Y_2 = \text{professional}|Y_1 = \text{farm}) &= \frac{\Pr(Y_2 = \text{professional} \cap Y_1 = \text{farm})}{\Pr(Y_1 = \text{farm})} \\ &= \frac{.018}{.018 + .035 + .031 + .008 + .018} \\ &= .164\,.\end{aligned}$$

In this example, we saw how to derive the conditional probabilities $p_{Y_2|Y_1}$ and $p_{Y_1|Y_2}$ from the joint probabilities p_{Y_1,Y_2}. Likewise we can also specify the joint from the marginal p_{Y_1} and the conditional $p_{Y_2|Y_1}$. In any case, we essentially need to specify 5×5 entries for the joint probability values $\Pr(Y_1 = y_1, Y_2 = y_2)$. Without further assumption, we needs $25 - 1 = 24$ parameters for the joint pdf, one for each probability value.

Suppose now that we wish to extend the joint pdf to describe the social mobility not for two but three or more generations. Assume that the list of occupations remain 5 in this example. With three generations (of grandfathers, fathers, sons) we need to specify $5^3 = 125$ entries for the joint pdf. With four generations, we need $5^4 = 625$ entries. And so on. This shows a fundamental challenge in working with multivariate domains. Without further assumptions, the number of parameters required is exponential in the number of variables. This would be unworkable. □

The main tool that statistical modelers exploit to overcome the complexity in modeling multivariate domains is to make use of independence, more appropriately, conditional independence, by incorporating our domain knowledge about the variables of interest.

Example 2.4. Continuing from the previous example. Let Y_1, Y_2, Y_3 denote the grandfather, father and son's occupations.

By chain rule, we may always write[2]

$$p(y_1, y_2, y_3) = p(y_1)p(y_2|y_1)p(y_3|y_1, y_2).$$

We may help ourselves by making the following assumption: *assume that Y_3 is conditionally independent of Y_1 given Y_2.*

This means, the joint conditional density of Y_1 and Y_3 given Y_2 equals the product the corresponding marginal conditional densities:

$$p(y_1, y_3|y_2) = p(y_1|y_2)p(y_3|y_2)$$

for any numerical values (y_1, y_2, y_3). The reader should verify that under the above conditional independence:

$$\begin{aligned} p(y_1|y_2, y_3) &= p(y_1|y_2) \\ p(y_3|y_2, y_1) &= p(y_3|y_2). \end{aligned}$$

As a consequence, we may specify the joint pdf of Y_1, Y_2, Y_3 by a smaller number of parameters, by noting that (why?)

$$p(y_1, y_2, y_3) = p(y_1)p(y_2|y_1)p(y_3|y_2).$$

Question: how many parameters do we need to specify the joint pdf?

Another question: suppose that the conditional distribution of the occupation of the grandfather generation given the father's is the same as the conditional distribution of that of the father's generation given the son's. How many parameters do we need now?

[2]Recall that we have removed the subscripts to avoid cluttering from $p_{Y_1,Y_2,Y_3}(y_1, y_2, y_3) = p_{Y_1}(y_1)p_{Y_2|Y_1}(y_2|y_1)p_{Y_3|Y_1,Y_2}(y_3|y_1, y_2)$.

2.5 Bayes' formulas and parameter estimation

As we described in Section 1, in order to initiate a Bayesian analysis, we need to specify the joint distribution of the quantity of interest θ and data y, by specifying the prior belief about θ via the prior distribution $p(\theta)$, and the sampling model $p(y|\theta)$. In practice, y represents the values of a collection of random variables/ vectors. θ is a random variable in a suitable domain. The principle of these specifications is the same as before, whether y and θ are discrete or continuous valued, or a combination thereof.

A large proportion of a Bayesian modeler's technical effort therefore is on finding a suitable specification of the joint distribution $p(\theta, y)$ for the problem at hand. Once this is done, having observed $\{Y = y\}$, we need to compute our updated beliefs about θ via the Bayes' formula, which is now expressed in terms of density function for random variables:

$$p(\theta|y) = p(\theta, y)/p(y) = p(\theta)p(y|\theta)/p(y). \tag{3}$$

Another significant proportion of the Bayesian framework is to compute that above posterior density function of θ, expressed above as a ratio.

- The numerator is the product between the prior pdf, $p(\theta)$, and the quantity $p(y|\theta)$.
- As a function of y, we call $p(y|\theta)$ the pdf of the sampling model, where θ plays the role of the parameter.
- As a function of θ, we call $p(y|\theta)$ as the *likelihood* function, with data y being fixed.

It's worth repeating that the likelihood function is *not* a density function. As the focus is shifted toward the inference of θ, "likelihood function" will be invoked more often.

Although the numerator of the posterior density is often simple to compute because the prior component and the likelihood component are typically explicitly specified, the denominator is typically difficult to compute explicitly. It can be seen that

$$p(y) = \int p(\theta, y)\mathrm{d}\theta = \int p(\theta)p(y|\theta)\mathrm{d}\theta$$

which involves taking integration (or summation) over the space of $\theta \in \Theta$. The integration typically does not admit an explicit form.

One may be interested in the relative posterior density, by comparing its value at different numerical values of interest. Let θ_a and θ_b be two such numerical values of θ, and take

$$\begin{aligned}\frac{p(\theta_a|y)}{p(\theta_b|y)} &= \frac{p(\theta_a)p(y|\theta_a)/p(y)}{p(\theta_b)p(y|\theta_b)/p(y)} \\ &= \frac{p(\theta_a)p(y|\theta_a)}{p(\theta_b)p(y|\theta_b)}\end{aligned}$$

In the above, the computation of the relative posterior density does not require the computation of $p(y)$, because $p(y)$ does not depend on specific value of θ. Accordingly, we often write

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

where $\propto$ is called "proportional" up to a normalizing constant to ensure that the left hand side is a value pdf for θ. The normalizing constant is precisely $p(y)$ in this case.

In words, we write

$$\boxed{\text{posterior} \propto \text{prior} \times \text{likelihood.}}$$

This captures succintly and beautifully the spirit of Bayesian inference: the posterior belief about the quantity of interest is obtained from two sources of information: the prior belief, and empirical observations (via the likelihood). Moreover, these two sources are combined explicitly via a multiplicative operation. As a function of the quantity interest, you may take this as an update to the prior belief via a reweighting operation, where the weights are provided by the likelihood function.

Finally, in practice, we are interested in various properties of the posterior density function $p(\theta|y)$, rather than the density function itself. This helps us express more precisely our belief about the true θ, because of the Bayesian "doctrine" that we usually do not know the exact truth; we can only calculate our belief about such truth. We have seen in the example of Section 1.3 various quantities of interest, including the posterior mean and posterior variance, posterior mode, posterior probability of tails, various quantiles and confidence regions with respect to the posterior distribution.

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.