likelihood function:

$$\begin{aligned}\Pr(y_1=0, y_2=0 \,|\, \theta=1) &= (0.5)(0.5) = 0.25\\ \Pr(y_1=0, y_2=0 \,|\, \theta=0) &= (1)(1) = 1.\end{aligned}$$

These expressions follow from the fact that if the woman is a carrier, then each of her sons will have a 50% chance of inheriting the gene and so being affected, whereas if she is not a carrier then there is a probability close to 1 that a son of hers will be unaffected. (In fact, there is a nonzero probability of being affected even if the mother is not a carrier, but this risk—the mutation rate—is small and can be ignored for this example.)

Posterior distribution. Bayes' rule can now be used to combine the information in the data with the prior probability; in particular, interest is likely to focus on the posterior probability that the woman is a carrier. Using y to denote the joint data (y_1, y_2), this is simply

$$\begin{aligned}\Pr(\theta=1|y) &= \frac{p(y|\theta=1)\Pr(\theta=1)}{p(y|\theta=1)\Pr(\theta=1) + p(y|\theta=0)\Pr(\theta=0)}\\ &= \frac{(0.25)(0.5)}{(0.25)(0.5)+(1.0)(0.5)} = \frac{0.125}{0.625} = 0.20.\end{aligned}$$

Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes' rule provides a formal mechanism for determining the extent of the correction. The results can also be described in terms of prior and posterior odds. The prior odds of the woman being a carrier are $0.5/0.5 = 1$. The likelihood ratio based on the information about her two unaffected sons is $0.25/1 = 0.25$, so the posterior odds are $1 \cdot 0.25 = 0.25$. Converting back to a probability, we obtain $0.25/(1+0.25) = 0.2$, as before.

Adding more data. A key aspect of Bayesian analysis is the ease with which sequential analyses can be performed. For example, suppose that the woman has a third son, who is also unaffected. The entire calculation does not need to be redone; rather we use the previous posterior distribution as the new prior distribution, to obtain:

$$\Pr(\theta=1|y_1, y_2, y_3) = \frac{(0.5)(0.20)}{(0.5)(0.20)+(1)(0.8)} = 0.111.$$

Alternatively, if we suppose that the third son is affected, it is easy to check that the posterior probability of the woman being a carrier becomes 1 (again ignoring the possibility of a mutation).

Spelling correction

Classification of words is a problem of managing uncertainty. For example, suppose someone types 'radom.' How should that be read? It could be a misspelling or mistyping of 'random' or 'radon' or some other alternative, or it could be the intentional typing of 'radom' (as in its first use in this paragraph). What is the probability that 'radom' actually means random? If we label y as the data and θ as the word that the person was intending to type, then

$$\Pr(\theta \,|\, y = \text{'radom'}) \propto p(\theta)\, \Pr(y = \text{'radom'} \,|\, \theta). \tag{1.6}$$

This product is the unnormalized posterior density. In this case, if for simplicity we consider only three possibilities for the intended word, θ (random, radon, or radom), we can compute the posterior probability of interest by first computing the unnormalized density for all three values of theta and then normalizing:

$$p(\text{random}|\text{'radom'}) = \frac{p(\theta_1)p(\text{'radom'}|\theta_1)}{\sum_{j=1}^{3} p(\theta_j)p(\text{'radom'}|\theta_j)},$$

where θ_1=random, θ_2=radon, and θ_3=radom. The prior probabilities $p(\theta_j)$ can most simply come from frequencies of these words in some large database, ideally one that is adapted to the problem at hand (for example, a database of recent student emails if the word in question is appearing in such a document). The likelihoods $p(y|\theta_j)$ can come from some modeling of spelling and typing errors, perhaps fit using some study in which people were followed up after writing emails to identify any questionable words.

Prior distribution. Without any other context, it makes sense to assign the prior probabilities $p(\theta_j)$ based on the relative frequencies of these three words in some databases. Here are probabilities supplied by researchers at Google:

θ	$p(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

Since we are considering only these possibilities, we could renormalize the three numbers to sum to 1 ($p(\text{random}) = \frac{760}{760+60.5+3.12}$, etc.) but there is no need, as the adjustment would merely be absorbed into the proportionality constant in (1.6).

Returning to the table above, we were surprised to see the probability of 'radom' in the corpus being as high as it was. We looked up the word in Wikipedia and found that it is a medium-sized city: home to 'the largest and best-attended air show in Poland ... also the popular unofficial name for a semiautomatic 9 mm Para pistol of Polish design ...' For the documents that we encounter, the relative probability of 'radom' seems much too high. If the probabilities above do not seem appropriate for our application, this implies that we have prior information or beliefs that have not yet been included in the model. We shall return to this point after first working out the model's implications for this example.

Likelihood. Here are some conditional probabilities from Google's model of spelling and typing errors:

θ	$p(\text{'radom'}\|\theta)$
random	0.00193
radon	0.000143
radom	0.975

We emphasize that this likelihood function is *not* a probability distribution. Rather, it is a set of conditional probabilities of a particular outcome ('radom') from three different probability distributions, corresponding to three different possibilities for the unknown parameter θ.

These particular values look reasonable enough—a 97% chance that this particular five-letter word will be typed correctly, a 0.2% chance of obtaining this character string by mistakenly dropping a letter from 'random,' and a much lower chance of obtaining it by mistyping the final letter of 'radon.' We have no strong intuition about these probabilities and will trust the Google engineers here.

Posterior distribution. We multiply the prior probability and the likelihood to get joint probabilities and then renormalize to get posterior probabilities:

θ	$p(\theta)p(\text{'radom'}\|\theta)$	$p(\theta\|\text{'radom'})$
random	1.47×10^{-7}	0.325
radon	8.65×10^{-10}	0.002
radom	3.04×10^{-7}	0.673

Thus, conditional on the model, the typed word 'radom' is about twice as likely to be correct as to be a typographical error for 'random,' and it is very unlikely to be a mistaken instance of 'radon.' A fuller analysis would include possibilities beyond these three words, but the basic idea is the same.

Decision making, model checking, and model improvement. We can envision two directions to go from here. The first approach is to accept the two-thirds probability that the word was typed correctly or even to simply declare 'radom' as correct on first pass. The second option would be to question this probability by saying, for example, that 'radom' looks like a typo and that the estimated probability of it being correct seems much too high.

When we dispute the claims of a posterior distribution, we are saying that the model does not fit the data or that we have additional prior information not included in the model so far. In this case, we are only examining one word so lack of fit is not the issue; thus a dispute over the posterior must correspond to a claim of additional information, either in the prior or the likelihood.

For this problem we have no particular grounds on which to criticize the likelihood. The prior probabilities, on the other hand, are highly context dependent. The word 'random' is of course highly frequent in our own writing on statistics, 'radon' occurs occasionally (see Section 9.4), while 'radom' was entirely new to us. Our surprise at the high probability of 'radom' represents additional knowledge relevant to our particular problem.

The model can be elaborated most immediately by including contextual information in the prior probabilities. For example, if the document under study is a statistics book, then it becomes more likely that the person intended to type 'random.' If we label x as the contextual information used by the model, the Bayesian calculation then becomes,

$$p(\theta|x, y) \propto p(\theta|x)p(y|\theta, x).$$

To first approximation, we can simplify that last term to $p(y|\theta)$, so that the probability of any particular error (that is, the probability of typing a particular string y given the intended word θ) does not depend on context. This is not a perfect assumption but could reduce the burden of modeling and computation.

The practical challenges in Bayesian inference involve setting up models to estimate all these probabilities from data. At that point, as shown above, Bayes' rule can be easily applied to determine the implications of the model for the problem at hand.

1.5 Probability as a measure of uncertainty

We have already used concepts such as probability density, and indeed we assume that the reader has a fair degree of familiarity with basic probability theory (although in Section 1.8 we provide a brief technical review of some probability calculations that often arise in Bayesian analysis). But since the uses of probability within a Bayesian framework are much broader than within non-Bayesian statistics, it is important to consider at least briefly the foundations of the concept of probability before considering more detailed statistical examples. We take for granted a common understanding on the part of the reader of the mathematical definition of probability: that probabilities are numerical quantities, defined on a set of 'outcomes,' that are nonnegative, additive over mutually exclusive outcomes, and sum to 1 over all possible mutually exclusive outcomes.

In Bayesian statistics, probability is used as the fundamental measure or yardstick of uncertainty. Within this paradigm, it is equally legitimate to discuss the probability of 'rain tomorrow' or of a Brazilian victory in the soccer World Cup as it is to discuss the probability that a coin toss will land heads. Hence, it becomes as natural to consider the probability that an unknown estimand lies in a particular range of values as it is to consider the probability that the mean of a random sample of 10 items from a known fixed population of size 100 will lie in a certain range. The first of these two probabilities is of more interest after data have been acquired whereas the second is more relevant beforehand. Bayesian methods enable statements to be made about the partial knowledge available (based on data) concerning some situation or 'state of nature' (unobservable or as yet unobserved) in