Inferring Probability Models from Data

9

One very useful way to draw conclusions from a dataset is to fit a probability model to that dataset. Once this is done, we can apply any of our procedures to the probability model to (for example) predict new data or estimate properties of future data. For example, you could flip a coin ten times and see five heads and five tails. To be able to estimate the probability of seeing heads on a future flip, you would need to fit a model. But once you had done so, you could also estimate the probability that five future flips would give you three heads and two tails, and so on.

The first step is to choose a model. I have described a relatively small subset of available probability models (really!), and so the choice of model will mostly be obvious. For example, we will use binomial or geometric models for coin flips, and so on. The procedures I describe, however, are quite general. They extend in principle to any model. Furthermore, they do not require that the model be right. Surprisingly, you can often extract quite useful information from a dataset by fitting a model that isn't the model that produced the dataset.

Once you have a model, you need to estimate values for its parameters. For example, if you use a binomial model for a coin flip, then you need to determine the probability a flip will come up heads. There are two kinds of procedure for estimating parameter values. Maximum likelihood finds the values of the parameters that make the observed data most likely. Bayesian inference produces a posterior probability distribution on the parameter values, and extracts information from that. Mostly, the description of these procedures is straightforward, but actually using them takes a little practice, so the sections in this chapter are mostly worked examples.

9.1 Estimating Model Parameters with Maximum Likelihood

Assume we have a dataset $\mathcal{D} = \{\mathbf{x}\}$, and a probability model we believe applies to that dataset. Generally, application logic suggests the type of model (i.e. normal probability density; Poisson probability; geometric probability; and so on). But usually, we do not know values for the parameters of the model—for example, the mean and standard deviation of a normal distribution; the intensity of a poisson distribution; and so on. Notice that this situation is unlike what we have seen to date. In Chap. 5, we assumed that we knew parameters, and could then use the model to assign a probability to a set of data items $\mathcal{D}$. Here we *know* the value of $\mathcal{D}$, but don't know the parameters. Our model will be better or worse depending on how well we choose the parameters. We need a strategy to estimate the parameters of a model from a sample dataset. Notice how each of the following examples fits this pattern. There is an important, and widespread, convention that I shall adhere to. Unknown parameters are widely referred to as θ (which could be a scalar, or a vector, or, if you're lucky, much more interesting than that).

Example 9.1 (Inferring p from Repeated Flips—Binomial) Imagine we flip a coin N times, and count the number of heads h. An appropriate probability model for a set of independent coin flips is the binomial model $P_b(h; N, \theta)$, where θ is the probability a flipped coin comes up heads (which was written $p(H)$ or p in the binomial model). But we do not know $p(H)$, which is the parameter. I wrote this as θ because we do not know it. We need a strategy to extract a value of θ from the data.

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_9

Example 9.2 (Inferring p from Repeated Flips—Geometric) Imagine we flip coin repeatedly until we see a head. The number of flips has the geometric distribution with parameter $p(H)$. In this case, the data is a sequence of T's with a final H from the coin flips. There are N flips (or terms) and the last flip is a head. We know that an appropriate probability model is the geometric distribution $P_g(N;\theta)$. But we do not know $p(H)$, which is the parameter I wrote as θ.

Example 9.3 (Inferring the Intensity of Spam—Poisson) It is reasonable to assume that the number of spam emails one gets in an hour has a Poisson distribution. But what is the intensity parameter, written λ in the definition, of that distribution? We could count the number of spam emails that arrive in each of a set of distinct hours, giving a dataset of counts $\mathcal{D}$. We need a strategy to wrestle an estimate of the intensity parameter, which I shall write θ because we don't know it, from this dataset.

Example 9.4 (Inferring the Mean and Standard Deviation of Normal Data) Imagine we know for some reason that our data is well described by a normal distribution. The missing parameters are now the mean and standard deviation of the normal distribution that best represents the data.

9.1.1 The Maximum Likelihood Principle

We have a dataset $\mathcal{D}$, a family of probability models $P(\mathcal{D}|\theta)$. We need a "reasonable" procedure to estimate a value of θ so that the resulting model describes our data well. A natural choice is the value of θ that makes the data observed "most probable". If we knew θ, then the probability of observing the data $\mathcal{D}$ would be $P(\mathcal{D}|\theta)$. We can construct an expression for $P(\mathcal{D}|\theta)$ using our model. Now we know $\mathcal{D}$, and we don't know θ, so the value of $P(\mathcal{D}|\theta)$ is a function of θ. This function is known as the **likelihood**.

Definition 9.1 (Likelihood) The function $P(\mathcal{D}|\theta)$, which is a function of θ, is known as the **likelihood** of the data D, and is often written $\mathcal{L}(\theta)$ (or $\mathcal{L}(\theta;\mathcal{D})$ if you want to remember that data is involved).

The maximum likelihood principle asks for a choice of θ such that the probability of observing the data you actually see, is maximised. This should strike you as being a reasonable choice.

Definition 9.2 (Maximum Likelihood Principle) The maximum likelihood principle chooses θ such that $\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$ is maximised, as a function of θ.

For the examples we work with, the data will be **independent and identically distributed** or **IID**. This means that each data item is an idependently obtained sample from the same probability distribution (see Sect. 4.3.1). In turn, this means that the likelihood is a product of terms, one for each data item, which we can write as

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = \prod_{i \in \text{dataset}} P(\mathbf{x}_i|\theta).$$

There are two, distinct, important concepts we must work with. One is the unknown parameter(s), which we will write θ. The other is the *estimate* of the value(s), which we will write $\hat{\theta}$. This estimate is the best we can do—it may not be the "true" value of the parameter, which we will likely never know.

Procedure 9.1 (Estimating with Maximum Likelihood) Given a dataset $\{\mathcal{D}\}$, and a model with unknown parameter(s) θ, compute an estimate $\hat{\theta}$ of the value of the parameters by constructing the likelihood of the data under the model

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta)$$

which in our case will always be

$$\mathcal{L}(\theta) = \prod_{i \in \text{dataset}} P(\mathbf{x}_i|\theta).$$

Now estimate $\hat{\theta}$ as

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, \mathcal{L}(\theta).$$

You should see Procedure 9.1 as a straightforward recipe, because that's what it is. It is a highly successful recipe. The main difficulty in applying the recipe is actually finding the maximum. The following sections show examples for important cases.

9.1.2 Binomial, Geometric and Multinomial Distributions

Worked example 9.1 (Inferring *p(H)* for a Coin from Flips Using a Binomial Model) In N independent coin flips, you observe k heads. Use the maximum likelihood principle to infer $p(H)$.

Solution The coin has $\theta = p(H)$, which is the unknown parameter. We know that an appropriate probability model is the binomial model $P_b(k; N, \theta)$. We have that

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = P_b(k; N, \theta)$$

$$= \binom{N}{k} \theta^k (1-\theta)^{(N-k)}$$

which is a function of θ—the unknown probability that a coin comes up heads; k and N are known. We must find the value of θ that maximizes this expression. Now the maximum occurs when

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0.$$

We have

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \binom{N}{k} \left(k\theta^{k-1}(1-\theta)^{(N-k)} - \theta^k (N-k)(1-\theta)^{(n-k-1)} \right)$$

and this is zero when

$$k\theta^{k-1}(1-\theta)^{(N-k)} = \theta^k (N-k)(1-\theta)^{(N-k-1)}$$

so the maximum occurs when

$$k(1-\theta) = \theta(N-k).$$

(continued)

This means the maximum likelihood estimate is

$$\hat{\theta} = \frac{k}{N}.$$

Worked example 9.1 produces a result that seems natural to most people, and it's very likely that you would guess this form without knowing the maximum likelihood principle. But now we have a procedure we can apply to other problems. Notice one quirk of the method that this example exposes. Generally, the method is more reliable with more data. If you flip a coin once, and get a T, the procedure will estimate $\hat{\theta} = 0$, which should strike you as a poor estimate.

Worked example 9.2 (Inferring *p(H)* from Coin Flips Using a Geometric Model). You flip a coin N times, stopping when you see a head. Use the maximum likelihood principle to infer $p(H)$ for the coin.

Solution The coin has $\theta = p(H)$, which is the unknown parameter. We know that an appropriate probability model is the geometric model $P_g(N; \theta)$. We have that

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = P_g(N; \theta) = (1-\theta)^{(N-1)}\theta$$

which is a function of θ—the unknown probability that a coin comes up heads; N is known. We must find the value of θ that maximizes this expression. Now the maximum occurs when

$$\begin{aligned}\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 \ &= ((1-\theta)^{(N-1)} \\ &\quad -(N-1)(1-\theta)^{(N-2)}\theta)\end{aligned}$$

So the maximum likelihood estimate is

$$\hat{\theta} = \frac{1}{N}.$$

Most people don't guess the estimate of Worked example 9.2, though it usually seems reasonable in retrospect. Obtaining the maximum of the likelihood can get interesting, as the following worked example suggests.

Worked example 9.3 (Inferring Die Probabilities from Multiple Rolls and a Multinomial Distribution). You throw a die N times, and see n_1 ones, ... and n_6 sixes. Write $p_1, \ldots, p_6$ for the probabilities that the die comes up one, ..., six. Use the maximum likelihood principle to estimate $p_1, \ldots, p_6$ for a multinomial model.

Solution The data are N, $n_1, \ldots, n_6$. The parameters are $\theta = (p_1, \ldots, p_6)$. $P(\mathcal{D}|\theta)$ comes from the multinomial distribution. In particular,

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = \frac{n!}{n_1! \ldots n_6!} p_1^{n_1} p_2^{n_2} \ldots p_6^{n_6}$$

which is a function of $\theta = (p_1, \ldots, p_6)$. Now we want to maximize this function by choice of θ. Notice that we could do this by simply making all p_i very large—but this omits a fact, which is that $p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$. So we substitute using $p_6 = 1 - p_1 - p_2 - p_3 - p_4 - p_5$ (there are other, neater, ways of dealing with this issue, but they take more background knowledge). At the maximum, we must have that for all i,

$$\frac{\partial \mathcal{L}(\theta)}{\partial p_i} = 0$$

(continued)

which means that, for p_i, we must have

$$n_i p_i^{(n_i-1)}(1-p_1-p_2-p_3-p_4-p_5)^{n_6} - p_i^{n_i} n_6(1-p_1-p_2-p_3-p_4-p_5)^{(n_6-1)} = 0$$

so that, for each p_i, we have

$$n_i(1-p_1-p_2-p_3-p_4-p_5) - n_6 p_i = 0$$

or

$$\frac{p_i}{1-p_1-p_2-p_3-p_4-p_5} = \frac{n_i}{n_6}.$$

You can check that this equation is solved by

$$\hat{\theta} = \frac{1}{(n_1+n_2+n_3+n_4+n_5+n_6)}(n_1, n_2, n_3, n_4, n_5, n_6)$$

9.1.3 Poisson and Normal Distributions

Maximizing the likelihood presents a problem. We usually need to take derivatives of products, which can quickly lead to quite unmanageable expressions. There is a straightforward cure. The logarithm is a monotonic function for non-negative numbers (i.e. if $x > 0$, $y > 0$, $x > y$, then $\log(x) > \log(y)$). This means that the values of θ that maximise the log-likelihood are the same as the values that maximise the likelihood. This observation allows us to transform a product into a sum, and the derivative of a sum is easy.

Definition 9.3 (Log-Likelihood of a Dataset Under a Model) The log-likelihood of a dataset under a model is a function of the unknown parameters, and you will often see it written as

$$\begin{aligned}\log \mathcal{L}(\theta) &= \log P(\mathcal{D}|\theta) \\ &= \sum_{i \in \text{dataset}} \log P(d_i|\theta).\end{aligned}$$

Worked example 9.4 (Poisson Distributions). You observe N intervals, each of the same, fixed length (in time, or space). You know that, in these intervals, events occur with a Poisson distribution (for example, you might be observing Prussian officers being kicked by horses, or telemarketer calls...). You know also that the intensity of the Poisson distribution is the same for each observation. The number of events you observe in the i'th interval is n_i. What is the intensity of the Poisson distribution?

Solution Write θ for the unknown intensity. The likelihood is

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i \in \text{intervals}} P(\{n_i \text{ events}\}|\theta) \\ &= \prod_{i \in \text{intervals}} \frac{\theta^{n_i} e^{-\theta}}{n_i!}.\end{aligned}$$

(continued)

It will be easier to work with logs. The log-likelihood is

$$\log \mathcal{L}(\theta) = \sum_i (n_i \log \theta - \theta - \log n_i!)$$

so that we must solve

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \sum_i (\frac{n_i}{\theta} - 1) = 0$$

which yields a maximum likelihood estimate of

$$\hat{\theta} = \frac{\sum_i n_i}{N}$$

Worked example 9.5 Worked example 9.5 (The Intensity of Swearing). A famously sweary politician gives a talk. You listen to the talk, and for each of 30 intervals 1 min long, you record the number of swearwords. You record this as a histogram (i.e. you count the number of intervals with zero swear words, with one, etc.). For the first 10 intervals, you see

No. of swear words	0	1	2	3	4
No. of intervals	5	2	2	1	0

and for the following 20 intervals, you see

No. of swear words	0	1	2	3	4
No. of intervals	9	5	3	2	1

Assume that the politician's use of swearwords is Poisson. What is the intensity using the first 10 intervals? the second 20 intervals? all the intervals? why are they different?

Solution Use the expression from Worked example 9.12 to find

$$\begin{aligned}\hat{\theta}_{10} &= \frac{\text{total number of swearwords}}{\text{number of intervals}} \\ &= \frac{9}{10}\end{aligned}$$

$$\begin{aligned}\hat{\theta}_{20} &= \frac{\text{total number of swearwords}}{\text{number of intervals}} \\ &= \frac{21}{20}\end{aligned}$$

$$\begin{aligned}\hat{\theta}_{30} &= \frac{\text{total number of swearwords}}{\text{number of intervals}} \\ &= \frac{30}{30}.\end{aligned}$$

These are different because the maximum likelihood estimate is an *estimate*—we can't expect to recover the exact value from a dataset. Notice, however, that the estimates are quite close.

Worked example 9.6 (The Mean of a Normal Distribution). Assume we have $x_1, \ldots, x_N$, and we wish to model these data with a normal distribution. Use the maximum likelihood principle to estimate the mean of that normal distribution.

Solution The likelihood of a set of data values under the normal distribution with unknown mean θ and standard deviation σ is

$$\begin{aligned}\mathcal{L}(\theta) &= P(x_1, \ldots x_N|\theta, \sigma) \\ &= P(x_1|\theta, \sigma)P(x_2|\theta, \sigma)\ldots P(x_N|\theta, \sigma) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)\end{aligned}$$

and this expression is a moderate nuisance to work with. The log of the likelihood is

$$\begin{aligned}\log \mathcal{L}(\theta) = &\left(\sum_{i=1}^{N} -\frac{(x_i-\theta)^2}{2\sigma^2}\right) \\ &+ \quad \text{term not depending on } \theta.\end{aligned}$$

We can find the maximum by differentiating wrt θ and setting to zero, which yields

$$\begin{aligned}\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} &= \sum_{i=1}^{N} \frac{2(x_i-\theta)}{2\sigma^2} \\ &= 0 \\ &= \frac{1}{\sigma^2}\left(\sum_{i=1}^{N} x_i - N\theta\right)\end{aligned}$$

so the maximum likelihood estimate is

$$\hat{\theta} = \frac{\sum_{i=1}^{N} x_i}{N}$$

which probably isn't all that surprising. Notice we did not have to pay attention to σ in this derivation—we did not assume it was known, it just doesn't do anything.

Remember this: *The mean of a dataset is the maximum likelihood estimate of the mean of a normal model fit to that dataset.*

Worked example 9.7 (The Standard Deviation of a Normal Distribution). Assume we have $x_1, \ldots, x_N$ which are data that can be modelled with a normal distribution. Use the maximum likelihood principle to estimate the standard deviation of that normal distribution.

Solution Now we have to write out the log of the likelihood in more detail. Write μ for the mean of the normal distribution and θ for the unknown standard deviation of the normal distribution. We get

(continued)

$$\log \mathcal{L}(\theta) = \left(\sum_{i=1}^{N} -\frac{(x_i - \mu)^2}{2\theta^2} \right) - N \log \theta$$
$$+ \text{Term not depending on } \theta$$

We can find the maximum by differentiating wrt θ and setting to zero, which yields

$$\frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} = \frac{-2}{\theta^3} \sum_{i=1}^{N} \frac{-(x_i - \mu)^2}{2} - \frac{N}{\theta} = 0$$

so the maximum likelihood estimate is

$$\hat{\theta} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

which probably isn't all that surprising, either.

Remember this: *The standard deviation of a dataset is the maximum likelihood estimate of the standard deviation of a normal model fit to that dataset.*

You should notice that one could maximize the likelihood of a normal distribution with respect to mean and standard deviation in one go (i.e. I could have done Worked examples 9.14 and 9.16 in one worked example, instead of two). I did this example in two parts because I felt it was more accessible that way; if you object, you're likely to be able to fill in the details yourself very easily.

Remember this: *If you have many data items and a probability model, you really should use maximum likelihood to estimate the parameters of the model.*

9.1.4 Confidence Intervals for Model Parameters

Assume we have a dataset $\mathcal{D} = \{\mathbf{x}\}$, and a probability model we believe applies to that dataset. We know how to estimate appropriate parameter values by maximizing the likelihood function $L(\theta)$. But we do not yet have a way to think about how accurately the data determines the best choice of parameter value. In particular, we should like to be able to construct a confidence interval for a parameter. This interval should capture our certainty in the value of the parameter. If a small change in the dataset would cause a large change in the parameter we recovered, then the interval should be large. But if quite extensive changes in the dataset will result in about the same recovered value, the interval should be small.

For maximum likelihood problems, it is hard to apply the reasoning of Sect. 6.2 directly. However, the underlying notion—the confidence interval represents an interval within which the population mean will lie for most samples—is naturally associated with repeated experiments. When the data is explained by a parametric probability model, we can use that model to produce other possible datasets. If we compute a maximum likelihood estimate $\hat{\theta}$ of the parameters of the model, we can draw IID samples *from the model*. We then look at the estimate from that new dataset; the spread of the estimates yields our confidence interval.

A $(1 - 2\alpha)$ confidence interval for a parameter is an interval $[c_\alpha, c_{(1-\alpha)}]$. We construct the interval such that, if we were to perform a very large number of repetitions of our original experiment then estimate a parameter value for each, c_α would be the α quantile and $c_{(1-\alpha)}$ would be the $(1 - \alpha)$ quantile of those parameter values. We interpret this to mean that, with confidence $(1 - 2\alpha)$, the correct value of our parameter lies in this interval. This definition isn't really watertight. How do we

perform a very large number of repetitions? If we don't, how can we tell how good the confidence interval is? Nonetheless, we can construct intervals that illustrate how sensitive our original inference is to the data that we have.

There is a natural, simulation based, algorithm for estimating confidence intervals. The algorithm should feel so natural to you that you may already have guessed what to do. First, we compute the maximum likelihood estimate of the parameters, $\hat{\theta}$. We assume this estimate is right, but need to see how our estimates of $\hat{\theta}$ would vary with different collections of data from our model with that parameter value. So we compute a collection of simulated datasets, $\mathcal{D}_i$, each the same size as the original dataset. We obtain these by simulating $P(\mathcal{D}|\hat{\theta})$. Next, we compute a maximum likelihood estimate for each of these simulated datasets, $\hat{\theta}_i$. Finally, we compute the relevant percentiles of these datasets. The result is our interval.

Procedure 9.2 (Estimating Confidence Intervals for Maximum Likelihood Estimates Using Simulation) Assume we have a dataset $\mathcal{D} = \{x\}$ of N items. We have a parametric model of this data $p(x|\theta)$, and write the likelihood $\mathcal{L}(\theta; \mathcal{D}) = P(\mathcal{D}|\theta)$. We construct $(1 - 2\alpha)$ confidence intervals using the following steps.

1. Compute the maximum likelihood estimate of the parameter, $\hat{\theta} = \underset{\theta}{\text{argmax}}\, \mathcal{L}(\theta; \mathcal{D})$.
2. Construct R simulated datasets $\mathcal{D}_i$, each consisting of N IID samples drawn from $p(x|\hat{\theta})$.
3. For each such dataset, compute $\hat{\theta}_i = \underset{\theta}{\text{argmax}}\, \mathcal{L}(\theta; \mathcal{D})$.
4. Obtain $c_\alpha(\hat{\theta}_i)$, the α'th quantile of the collection $\hat{\theta}_i$ and $c_{(1-\alpha)}(\hat{\theta}_i)$, the $1 - \alpha$'th quantile of the collection $\hat{\theta}_i$.

The confidence interval is $[c_\alpha, c_{(1-\alpha)}]$.

Figure 9.1 shows an example. In this case, I worked with simulated data from a normal distribution. In each case, the normal distribution had mean 0, but there are two different standard deviations (1 and 10). I simulated 10 different datasets from each of these distributions, containing $10, 40, 90, \ldots, 810, 1000$ data items. For each, I computed the maximum likelihood estimate of the mean. This isn't zero, *even though the data was drawn from a zero-mean distribution*, because the dataset is finite. I then estimated the confidence intervals for each using 10,000 simulated datasets of the same size. I show 95% confidence intervals for the two cases, plotted against the size of the dataset used for the estimate. Notice that these intervals aren't symmetric about zero, because the maximum likelihood estimate isn't zero. They shrink as the dataset grows, but slowly. They are bigger when the standard deviation is bigger. It should seem reasonable that you can't expect an accurate estimate of the mean of a normal distribution with large standard deviation using only a few data points. If you think of a probability model as being like a population (it's a very large dataset, that is hidden, and from which you draw samples with replacement), then it should also seem reasonable to you that the intervals shrink as N goes up, because this is what the reasoning of Sect. 6.2.2 predicts.

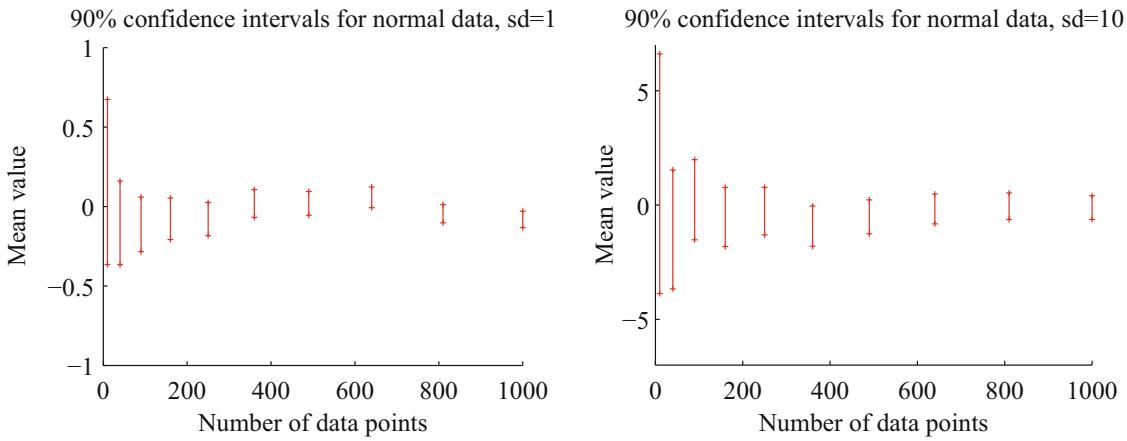

Fig. 9.1 Confidence intervals computed for simulated normal data; details in the text

Worked example 9.8 (Confidence Intervals by Simulation—II). Construct a 90% confidence interval for the intensity estimate for the data of Example 9.5 for the cases of 10 observations, 20 observations, and all 30 observations.

Solution Recall from that example the maximum likelihood estimates of the intensity are 7/10, 22/20, and 29/30 in the three cases. I used the Matlab function `poissrnd` to get 10,000 replicates of a dataset of 10 (resp. 20, 30) items from a Poisson distribution with the relevant intensities. I then used `prctile` to get the 5% and 95% percentiles, yielding the intervals

[0.3, 1.2]	For 10 observations
[0.75, 1.5]	For 20 observations
[0.6667, 1.2667]	For 30 observations

Notice how having more observations makes the confidence interval smaller.

9.1.5 Cautions About Maximum Likelihood

The maximum likelihood principle has a variety of neat properties we cannot expound. One worth knowing about is **consistency**; for our purposes, this means that the maximum likelihood estimate of parameters can be made arbitrarily close to the right answer by having a sufficiently large dataset. Now assume that our data doesn't actually come from the underlying model. This is the usual case, because we can't usually be sure that, say, the data truly is normal or truly comes from a Poisson distribution. Instead we choose a model that we think will be useful. When the data doesn't come from the model, maximum likelihood produces an estimate of θ that corresponds to the model that is (in quite a strong sense, which we can't explore here) the closest to the source of the data. Maximum likelihood is very widely used because of these neat properties. But there are some problems. One important problem is that it might be hard to find the maximum of the likelihood exactly. There are strong numerical methods for maximizing functions, and these are very helpful, but even today there are likelihood functions where it is very hard to find the maximum.

The second is that small amounts of data can present nasty problems. For example, in the binomial case, if we have only one flip we will estimate p as either 1 or 0. We should find this report unconvincing. In the geometric case, with a fair coin, there is a probability 0.5 that we will perform the estimate and then report that the coin has $p = 1$. This should also worry you. As another example, if we throw a die only a few times, we could reasonably expect that, for some i, $n_i = 0$. This doesn't necessarily mean that $p_i = 0$, though that's what the maximum likelihood inference procedure will tell us.

This creates a very important technical problem—how can I estimate the probability of events that haven't occurred? This might seem like a slightly silly question to you, but it isn't. Solving this problem has really significant practical consequences. As one example, consider a biologist trying to count the number of butterfly species on an island. The biologist catches and classifies a lot of butterflies, then leaves. But are there more butterfly species on the island? To get some sense that we can reason successfully about this problem, compare two cases. In the first, the biologist catches many individuals of each of the species observed. In this case, you should suspect that catching more butterflies is unlikely to yield more species. In the second case, there are many species where the biologist sees only one individual of that species. In this case, you should suspect that catching more butterflies might very well yield new species.

9.2 Incorporating Priors with Bayesian Inference

Another important issue with maximum likelihood is that there is no mechanism to incorporate prior beliefs. For example, imagine you get a new die from a reliable store, roll it six times and see a one once. You would be happy to believe that $p(6) = 1/6$ for this die. Now imagine you borrow a die from a friend with a long history of making weighted dice. Your friend tells you this die is weighted so that $p(1) = 1/2$. You roll the die six times and see a one once; in this case, you might worry that $p(1)$ isn't $1/6$, and you just happened to get a slightly unusual set of rolls. You'd worry because you have

good reason to believe the die isn't fair, and you'd want more evidence to believe $p(6) = 1/6$. Maximum likelihood can't distinguish between these two cases.

The difference lies in prior information—information we possess before we look at the data. We would like to take this information into account when we estimate the model. One way to do so is to place a **prior probability distribution** $p(\theta)$ on the parameters θ. Then, rather than working with the likelihood $p(\mathcal{D}|\theta)$, we could apply Bayes' rule, and form the **posterior** $P(\theta|\mathcal{D})$. This posterior represents the probability that θ takes various values, given the data $\mathcal{D}$. Bayes' rule tells us that

$$\begin{aligned} P(\theta|\mathcal{D}) &= \frac{P(\mathcal{D}|\theta) \times P(\theta)}{P(\mathcal{D})} \\ &= \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalizing constant}}. \end{aligned}$$

Be aware that the prior distribution is often written with a π rather than a P or a p.

Definition 9.4 (Bayesian Inference) Extracting information from the posterior $P(\theta|\mathcal{D})$ is usually called **Bayesian inference**.

Having the posterior probability distribution immediately allows us to answer quite sophisticated questions. For example, we could immediately compute

$$P(\{\theta \in [0.2, 0.4]\}|\mathcal{D})$$

in a straightforward way. Quite often, we just want to extract an estimate of θ. For this, we can use the θ that maximizes the posterior.

Definition 9.5 (MAP Estimate) A natural estimate of θ is the value that maximizes the posterior $P(\theta|\mathcal{D})$. This estimate is known as a **maximum a posteriori estimate** or **MAP estimate**.

To get this θ, we do not need to know the *value* of the posterior, and it is enough to work with

$$P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta).$$

$$\begin{aligned} P(\theta|\mathcal{D}) &\propto P(\mathcal{D}|\theta) \times P(\theta) \\ &\propto \text{Likelihood} \times \text{Prior} \end{aligned}$$

This form exposes similarities and differences between Bayesian and maximum likelihood inference. To reason about the posterior, you need to have a prior $P(\theta)$. If you assume that this prior has the same value for every θ, you'll end up doing maximum likelihood, because $P(\theta)$ is some constant. Using a prior that isn't constant means that the MAP estimate may be different from the maximum likelihood estimate. This difference is occurring because you have prior beliefs that some θ are more likely to occur than others. Supplying these prior beliefs is part of the modelling process.

Worked example 9.9 (Flipping a Coin). We have a coin with probability θ of coming up heads when flipped. We start knowing nothing about θ. We then flip the coin 10 times, and see 7 heads (and 3 tails). Plot a function proportional to $p(\theta|\{7 \text{ heads and } 3 \text{ tails}\})$. What happens if there are 3 heads and 7 tails?

Solution We know nothing about p, except that $0 \leq \theta \leq 1$, so we choose a uniform prior on p. We have that $p(\{7 \text{ heads and } 3 \text{ tails}\}|\theta)$ is binomial. The joint distribution is $p(\{7 \text{ heads and } 3 \text{ tails}\}|\theta) \times p(\theta)$ but $p(\theta)$ is uniform, so doesn't depend on θ. So the posterior is *proportional* to: $\theta^7(1-\theta)^3$ which is graphed in Fig. 9.2. The figure also

(continued)

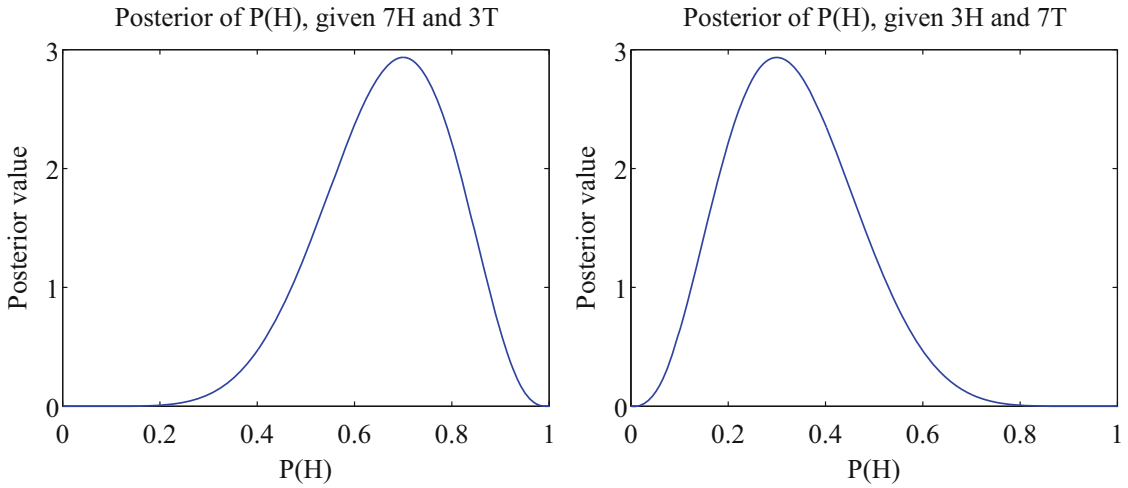

Fig. 9.2 The curves show a function *proportional to* the posterior on θ, for the two cases of Example 9.9. Notice that this information is rather richer than the single value we would get from maximum likelihood inference

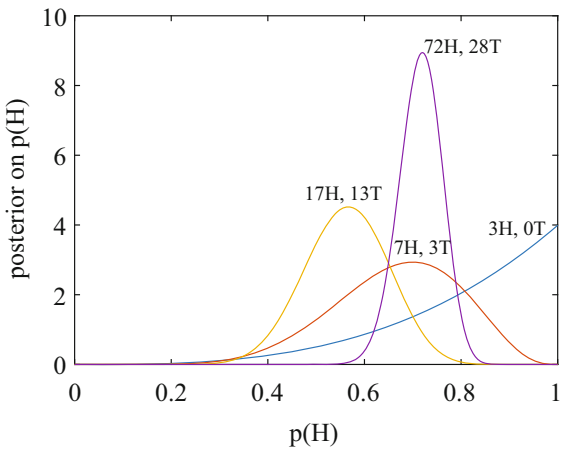

Fig. 9.3 The probability that an unknown coin will come up heads when flipped is $p(H)$. For these figures, I simulated coin flips from a coin with $p = 0.75$. I then plotted the posterior for various data. Notice how, as we see more flips, we get more confident about p. The graph gets higher as it gets narrower because the posterior probability must integrate to one

shows $\theta^3(1-\theta)^7$ which is *proportional* to the posterior for 3 heads and 7 tails. In each case, the evidence does not rule out the possibility that $\theta = 0.5$, but tends to discourage the conclusion. Maximum likelihood would give $\theta = 0.7$ or $\theta = 0.3$, respectively.

Figure 9.2 shows curves *proportional to* the posterior. The functions are not the true posterior, because they do not integrate to one. The fact that we are missing this constant of proportionality means that, for example, we cannot compute $P(\{\theta \in [0.3, 0.6]\} | \mathcal{D})$. The constant of proportionality can be computed by noticing that

$$P(\mathcal{D}) = \int_{\theta} P(\mathcal{D}|\theta)P(\theta)d\theta.$$

It is usually impossible to do this integral in closed form, so we would have to use a numerical integral or a trick. For the case of Fig. 9.2, it is fairly easy to estimate the constant of proportionality (using either numerical integration or the fact that they're proportional to a binomial distribution). Figure 9.3 shows a set of true posteriors for different sets of evidence, using a uniform prior. The integral becomes much harder for more elaborate problems.

9.2.1 Conjugacy

Here is one really useful trick for computing the normalizing constant. In some cases, $P(\theta)$ and $P(\mathcal{D}|\theta)$, when multiplied together, take a familiar form. This happens when $P(\mathcal{D}|\theta)$ and $P(\theta)$ each belong to parametric families where there is a special relationship between the families. The property is known as **conjugacy**, and when a prior has this property, it is called a **conjugate prior**. The property is best captured in examples; the two in this section are occasionally worth knowing, but the most important example is in a section on its own (section Worked example 14.13).

Worked example 9.10 (Flipping a Coin—II). We have a coin with probability θ of coming up heads when flipped. We model the prior on θ with a Beta distribution, with parameters $\alpha > 0$, $\beta > 0$. We then flip the coin N times, and see h heads. What is $P(\theta|N, h, \alpha, \beta)$?

Solution We have that $P(N, h|\theta)$ is binomial, and that $P(\theta|N, h, \alpha, \beta) \propto P(N, h|\theta)P(\theta|\alpha, \beta)$. This means that

$$P(\theta|N, h, \alpha, \beta) \propto \binom{N}{h} \theta^h (1-\theta)^{(N-h)}$$

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}.$$

and we can write

$$P(\theta|N, h, \alpha, \beta) \propto \theta^{(\alpha+h-1)}(1-\theta)^{(\beta+N-h-1)}.$$

Notice this has the form of a Beta distribution, so it is easy to recover the constant of proportionality. We have

$$P(\theta|N, h, \alpha, \beta) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)} \theta^{(\alpha+h-1)}(1-\theta)^{(\beta+N-h-1)}.$$

The normalizing constant for $P(\theta|N, h, \alpha, \beta)$ was easy to recover because the Beta distribution is a conjugate prior for the binomial distribution.

Remember this: *The Beta distribution is a conjugate prior for the binomial distribution*

Worked example 9.11 (More Sweary Politicians). Example 9.5 gives some data from a sweary politician. Assume we have only the first 10 intervals of observations, and we wish to estimate the intensity using a Poisson model. Write θ for this parameter. Use a Gamma distribution as a prior, and write out the posterior.

Solution We have that

$$\begin{aligned} p(\mathcal{D}|\theta) &= \left(\frac{\theta^0 e^{-\theta}}{0!}\right)^5 \left(\frac{\theta^1 e^{-\theta}}{1!}\right)^2 \times \\ &\quad \left(\frac{\theta^2 e^{-\theta}}{2!}\right)^2 \left(\frac{\theta^3 e^{-\theta}}{3!}\right)^1 \\ &= \frac{\theta^9 e^{-10\theta}}{12} \end{aligned}$$

(continued)

and

$$p(\theta|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{(\alpha-1)}e^{-\beta\theta}$$

This means that

$$p(\theta|\mathcal{D},\alpha,\beta) \propto \theta^{(\alpha-1+9)}e^{-(\beta+10)\theta}.$$

Notice this has the form of another Gamma distribution, so we can write

$$p(\theta|\mathcal{D},\alpha,\beta) = \frac{(\beta+10)^{(\alpha+9)}}{\Gamma(\alpha+9)}\theta^{(\alpha-1+9)}e^{-(\beta+10)\theta}$$

The normalizing constant for $P(\theta|\mathcal{D},\alpha,\beta)$ was easy to recover because the Gamma distribution is a conjugate prior for the Poisson distribution.

Remember this: *The Gamma distribution is a conjugate prior for the Poisson distribution*

9.2.2 MAP Inference

Look at Example 9.1, where we estimated the probability a coin would come up heads with maximum likelihood. We could not change our estimate just by knowing the coin was fair, but we could come up with a number for $\theta = p(H)$ (rather than, say, a posterior distribution). A natural way to produce a point estimate for θ that incorporates prior information is to choose $\hat{\theta}$ such that

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, P(\theta|\mathcal{D}) = \underset{\theta}{\text{argmax}}\, \frac{P(\theta,\mathcal{D})}{P(\mathcal{D})}$$

This is the MAP estimate. If we wish to perform MAP inference, $P(\mathcal{D})$ doesn't matter (it changes the value, but not the location, of the maximum). This means we can work with $P(\theta,\mathcal{D})$, often called the **joint** distribution.

Worked example 9.12 (Flipping a Coin—II). We have a coin with probability θ of coming up heads when flipped. We model the prior on θ with a Beta distribution, with parameters $\alpha > 0$, $\beta > 0$. We then flip the coin N times, and see h heads. What is the MAP estimate of θ?

Solution We have that

$$P(\theta|N,h,\alpha,\beta) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+h)\Gamma(\beta+N-h)}\theta^{(\alpha+h-1)}(1-\theta)^{(\beta+N-h-1)}.$$

You can get the MAP estimate by differentiating and setting to 0, yielding

$$\hat{\theta} = \frac{\alpha-1+h}{\alpha+\beta-2+N}.$$

This has rather a nice interpretation. You can see α and β as extra counts of heads (resp. tails) that are added to the observed counts. So, for example, if you were fairly sure that the coin should be fair, you might make α and β large and equal. When $\alpha = 1$ and $\beta = 1$, we have a uniform prior as in the previous examples.

Worked example 9.13 Worked example 9.13 (More Sweary Politicians). We observe our swearing politician for N intervals, seeing n_i swear words in the i'th interval. We model the swearing with a Poisson model. We wish to estimate the intensity, which we write θ. We use a Gamma distribution for the prior on θ. What is the MAP estimate of θ?

Solution Write $T = \sum_{i=1}^{N}$. We have that

$$p(\theta|\mathcal{D}) = \frac{(\beta + N)^{(\alpha+T)}}{\Gamma(\alpha + T)}\theta^{(\alpha-1+T)}e^{-(\beta+T)\theta}$$

and the MAP estimate is

$$\hat{\theta} = \frac{(\alpha - 1 + T)}{(\beta + N)}$$

(which you can get by differentiating with respect to θ, then setting to zero). Notice that if β is close to zero, you can interpret α as extra counts; if β is large, then it strongly discourages large values of $\hat{\theta}$, even if the counts are large.

Useful Facts 9.1 (Bayesian Inference is Particularly Good with Little Data)
If you have few data items and a model and a reasonable choice of prior, you really should use Bayesian inference.

9.2.3 Cautions About Bayesian Inference

Just like maximum likelihood inference, bayesian inference is not a recipe that can be applied without thought. It turns out that, when there is a lot of data, the prior has little influence on the outcome of the inference, and the MAP solution looks a lot like the maximum likelihood solution. So the difference between the two approaches is most interesting when there is little data, where the prior matters. The difficulty is that it might be hard to know what to use as a good prior. In the examples, I emphasized mathematical convenience, choosing priors that lead to clean posteriors. There is no reason to believe that nature uses conjugate priors (even though conjugacy is a neat property). How should one choose a prior for a real problem?

This isn't an easy point. If there is little data, then the choice could really affect the inference. Sometimes we're lucky, and the logic of the problem dictates a choice of prior. Mostly, we have to choose and live with the consequences of the choice. Often, doing so is successful in applications.

The fact we can't necessarily justify a choice of prior seems to be one of life's inconveniences, but it represents a significant philosophical problem. It's been at the core of a long series of protracted, often quite intense, arguments about the philosophical basis of statistics. I haven't followed these arguments closely enough to summarize them; they seem to have largely died down without any particular consensus being reached.

9.3 Bayesian Inference for Normal Distributions

Normal distribution models allow a variety of quite special tricks. Some algebra will establish that a normal prior and a normal likelihood yield a normal posterior. This turns out to be quite useful, because representing normal distributions is easy, and we can write straightforward equations for the mean and standard deviation of the posterior given prior and likelihood. You should remember from Sect. 1.3.3 that it is possible to compute the mean and standard deviation of a dataset if you see it one element at a time. Remarkably, the rules for can be put into a similar form. This means that bayesian inference for some kinds of time signal is quite straightforward.

9.3.1 Example: Measuring Depth of a Borehole

Assume we drop a measuring device down a borehole. A braking system will stop its fall and catch onto the side of the hole after it has fallen μ_π meters. On board is a device to measure its depth. The measuring device measures depth in feet (*not* meters: it's my example, so I can do what I like *and* I've seen this sort of foolishness in real life). This device reports the correct depth in feet plus a zero mean normal random variable, which we call "noise". The device reports depth every second, over wireless.

The first question to ask is what depth do we believe the device is at *before* we receive any measurement? We designed the braking system to stop at μ_π meters, so we are not completely ignorant about where it is. However, it may not have worked absolutely correctly. We choose to model the depth at which it stops as μ_π meters plus a zero mean normal random variable ("noise"). The noise term could be caused by error in the braking system, etc. We could estimate the standard deviation of the noise term (which we write σ_π) either by dropping devices down holes, then measuring with tape measures, or by analysis of likely errors in our braking system. The depth of the object is the unknown parameter of the model; we write this depth θ. Now the model says that θ is a normal random variable with mean μ_π and standard deviation σ_π.

Now assume we receive a single measurement—what do we now know about the device's depth? The first thing to notice is that there is something to do here. Ignoring the prior and taking the measurement might not be wise. For example, imagine that the noise in the wireless system is large, so that the measurement is often corrupted. In this case, our original guess about the device's location might be better than the measurement. Similarly, ignoring the measurement and just taking the prior is unwise, too. The measurement tells us something about the borehole that isn't represented in the prior, and we should use that.

Another reason to use the measurement is that we will receive another measurement in a second's time. Remember that each measurement is the true depth in feet plus zero-mean noise. Averaging multiple measurements will produce an estimate of depth that improves as the number of measurements goes up (by the standard error reasoning of Sect. 6.2.2). Since measurements will keep arriving, we want an online procedure that gives the best estimate of depth with current measurements, and then updates that estimate when a new measurement arrives. It turns out such a procedure is easy to construct.

9.3.2 Normal Prior and Normal Likelihood Yield Normal Posterior

When both $P(\mathcal{D}|\theta)$ and $P(\theta)$ are normal with known standard deviation, the posterior is normal, too. The mean and standard deviation of the posterior take a simple form. Assume $P(\theta)$ is normal, with mean μ_π and standard deviation σ_π (remember, priors are often written with π somewhere). So

$$\log P(\theta) = -\frac{(\theta - \mu_\pi)^2}{2\sigma_\pi^2} + \text{constant not dependent on } \theta.$$

Start by assuming that $\mathcal{D}$ is a single measurement x_1. The measurement x_1 could be in different units from θ, and we will assume that the relevant scaling constant c_1 is known. We assume that $P(x_1|\theta)$ is normal with known standard deviation $\sigma_{m,1}$, and with mean $c_1\theta$. Equivalently, x_1 is obtained by adding noise to $c_1\theta$. The noise will have zero mean and standard deviation $\sigma_{m,1}$. This means that

$$\log P(\mathcal{D}|\theta) = \log P(x_1|\theta) = -\frac{(x_1 - c_1\theta)^2}{2\sigma_{m,1}^2} + \text{constant not dependent on } x_1 \text{ or } \theta.$$

We would like to know $P(\theta|x)$. We have that

$$\log P(\theta|x_1) = \log p(x_1|\theta) + \log p(\theta) + \text{terms not depending on } \theta$$

$$
\begin{aligned}
&= -\frac{(x_1 - c_1\theta)^2}{2\sigma_{m,1}^2} - \frac{(\theta - \mu_\pi)^2}{2\sigma_\pi^2} \\
&\quad + \text{terms not depending on } \theta. \\
&= -\left[\theta^2\left(\frac{c_1^2}{2\sigma_{m,1}^2} + \frac{1}{2\sigma_\pi^2}\right)\right. \\
&\quad \left. -\theta\left(\frac{c_1 x_1}{2\sigma_{m,1}^2} + \frac{\mu_\pi}{2\sigma_\pi^2}\right)\right] \\
&\quad + \text{terms not depending on } \theta.
\end{aligned}
$$

Now some trickery will get us an expression for $P(\theta|x_1)$. Notice first that $\log P(\theta|x_1)$ is of degree 2 in θ (i.e. it has terms θ^2, θ and things that don't depend on θ). This means that $P(\theta|x_1)$ must be a normal distribution, because we can rearrange its log into the form of the log of a normal distribution.

Now we can show that $P(\theta|\mathcal{D})$ is normal when there are more measurements. Assume we have N measurements, $x_1, \ldots, x_N$. The measurements are IID samples from a normal distribution conditioned on θ. We will assume that each measurement is in its own set of units (captured by a constant c_i), and each measurement incorporates noise of different standard deviation (with standard deviation $\sigma_{m,i}$). So

$$
\begin{aligned}
\log P(x_i|\theta) &= -\frac{(x_i - c_i\theta)^2}{2\sigma_{m,i}^2} \\
&\quad + \text{constant not dependent on } x_1 \text{ or } \theta.
\end{aligned}
$$

Now

$$
\log P(\mathcal{D}|\theta) = \sum_i \log P(x_i|\theta)
$$

so we can write

$$
\begin{aligned}
\log P(\theta|\mathcal{D}) &= \log p(x_N|\theta) + \ldots + \log p(x_2|\theta) \\
&\quad + \log p(x_1|\theta) + \log p(\theta) \\
&\quad + \text{terms not depending on } \theta \\
&= \log p(x_N|\theta) + \ldots + \log p(x_2|\theta) \\
&\quad + \log p(\theta|x_1) + \text{terms not} \\
&\quad \text{depending on } \theta \\
&= \log p(x_N|\theta) + \ldots + \log p(\theta|x_1, x_2) \\
&\quad + \text{terms not depending on } \theta.
\end{aligned}
$$

This lays out the induction. We have that $P(\theta|x_1)$ is normal, with known standard deviation. Now regard this as the prior, and $P(x_2|\theta)$ as the likelihood; we have that $P(\theta|x_1, x_2)$ is normal, and so on. So under our assumptions, $P(\theta|\mathcal{D})$ is normal. We now have a really useful fact.

Remember this: *A normal prior and a normal likelihood yield a normal posterior when both standard deviations are known*

Some straightforward thrashing through algebra, relegated to the exercises, will yield expressions for the mean and standard deviation of the posterior. These are captured in the box below.

Useful Facts 9.2 (The Parameters of a Normal Posterior with a Single Measurement)
Assume we wish to estimate a parameter θ. The prior distribution for θ is normal, with known mean μ_π and known standard deviation σ_π. We receive a single data item x_1 and a scale c_1. The likelihood of x_1 is normal with mean $c_1\theta$ and standard deviation $\sigma_{m,1}$, where $\sigma_{m,1}$ is known. Then the posterior, $p(\theta|x_1, c_1, \sigma_{m,1}, \mu_\pi, \sigma_\pi)$, is normal, with mean

$$\mu_1 = \frac{c_1 x_1 \sigma_\pi^2 + \mu_\pi \sigma_{m,1}^2}{\sigma_{m,1}^2 + c_1^2 \sigma_\pi^2}$$

and standard deviation

$$\sigma_1 = \sqrt{\frac{\sigma_{m,1}^2 \sigma_\pi^2}{\sigma_{m,1}^2 + c^2 \sigma_\pi^2}}.$$

The equations of box 9.2 "make sense". Recall the example of dropping the depth measuring device down a borehole. Imagine that the mechanical design for the braking system was very good, but the measuring system is inaccurate. Then σ_π is very small, and $\sigma_{m,1}$ is very big. In turn, the mean of $P(\theta|x_1)$ is about μ_π. Equivalently, because our prior was very accurate, and the measurement was unreliable, the posterior mean is about the prior mean Similarly, if the measurement is reliable (i.e. $\sigma_{m,1}$ is small) and the prior has high variance (i.e. σ_π is large), the mean of $P(\theta|x_1)$ is about x_1/c_1.—i.e. the measurement, rescaled to the same units as θ

Worked example 9.14 (MAP for Normal Prior and Likelihood with Known Standard Deviation). We wish to estimate a parameter θ. The prior distribution for θ is normal, with known mean μ_π and known standard deviation σ_π. We have a single data item x_1. The likelihood $P(x_1|\theta)$ is normal, with mean $c_1\theta$ and standard deviation $\sigma_{m,1}$. What is the MAP estimate of θ?

Solution The equations are in a box, above (page 214). A normal distribution has a maximum at the mean, so

$$\hat{\theta} = \frac{c_1 x_1 \sigma_\pi^2 + \mu_\pi \sigma_{m,1}^2}{\sigma_{m,1}^2 + c_1^2 \sigma_\pi^2}$$

9.3.3 Filtering

Recall the device we dropped down a borehole produced a measurement every second. It does not make sense to wait all day, then use the day's measurements to produce a single depth estimate. Instead, we should update our estimate each time we get a measurement. The induction sketched in Sect. 9.3.2 gives a procedure to do this.

In words, our initial representation of the parameters we are trying to estimate is the prior, which is normal. We see one measurement, which has normal likelihood, so the posterior is normal. You can think of this posterior as a prior for the parameter estimate based on the *next* measurement. But we know what to do with a normal prior, a normal likelihood, and a measurement, so we can incorporate the measurement and go again. This means we can exploit our expression for the posterior mean and standard deviation in the case of normal likelihood and normal prior and a single measurement to deal with multiple measurements very easily. This process of updating a representation of a dataset as new data arrives is known as **filtering**.

To restate using mathematical notation, I will use the notation and assumptions of that section (so you have to read it, sorry). We assumed that the prior, $P(\theta)$, was normal, and the likelihood for the first measurement, $P(x_1|\theta)$, was normal. This meant that the posterior after the first measurement, $P(\theta|x_1)$, was normal too. Now we can treat $P(\theta|x_1, \ldots, x_{i-1})$ (which is

normal) as a prior for the likelihood $P(x_i|\theta)$ (which is also normal). This will produce the posterior $P(\theta|x_1, \ldots, x_i)$, which will be normal. This can operate as a prior for the likelihood $P(x_{i+1}|\theta)$, and so on. In turn, this gives us a pattern for incorporating measurements one at a time.

Useful Facts 9.3 (Normal Posteriors Can Be Updated Online)
Assume we wish to estimate a parameter θ. The prior distribution for θ is normal, with known mean μ_π and known standard deviation σ_π. We write x_i for the i'th data item. The likelihood for each separate data item is normal, with mean $c_i\theta$ and standard deviation $\sigma_{m,i}$. We have already received k data items. The posterior $p(\theta|x_1, \ldots, x_k, c_1, \ldots, c_k, \sigma_{m,1}, \ldots, \sigma_{m,k}, \mu_\pi, \sigma_\pi)$ is normal, with mean μ_k and standard deviation σ_k. We receive a new data item x_{k+1}. The likelihood of this data item is normal with mean $c_{k+1}\theta$ and standard deviation $\sigma_{m,(k+1)}$, where c_{k+1} and $\sigma_{m,(k+1)}$ are known. Then the posterior, $p(\theta|x_1, \ldots, x_{k+1}, c_1, \ldots, c_k, c_{k+1}, \sigma_{m,1}, \ldots, \sigma_{m,(k+1)}, \mu_\pi, \sigma_\pi)$, is normal, with mean

$$\mu_{k+1} = \frac{c_{k+1}x_{k+1}\sigma_k^2 + \mu_k\sigma_{m,(k+1)}^2}{\sigma_{m,(k+1)}^2 + c_{k+1}^2\sigma_k^2}$$

and

$$\sigma_{k+1}^2 = \frac{\sigma_{m,(k+1)}^2\sigma_k^2}{\sigma_{m,(k+1)}^2 + c_{k+1}^2\sigma_k^2}.$$

Again, notice the very useful fact that, if everything is normal, we can update our posterior representation when new data arrives using a very simple recursive form.

Worked example 9.15 (Estimating the Weekly Percent Growth in a Stock Price). Assume the weekly percent growth in the price of MSFT is an (unknown) constant. Use the stock price dataset at http://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index to estimate this constant at the end of each week. Assume that $\sigma_{m,i} = 1.9$ and that $c_i = 1$ for each i. Use $\mu_\pi = 0$ and $\sigma_\pi = 5$. Plot the mean and standard deviation of your current posterior at the end of each week. How do the standard deviations of the posteriors change with more measurements?

Solution This dataset is part of the UC Irvine Machine Learning archive. It was collected by Michael Brown. It gives a variety of numbers for a variety of stocks at the end of each of 25 weeks. The exercise requires a few lines of code to implement the recursions of box 9.3, and rather more lines of code to produce a plot. A little algebra will show that the k'th standard deviation should be proportional to $1/k$ (quick experiment shows the relevant constant is close to 1.7). Figure 9.4 shows the results.

9.4 You Should

9.4.1 Remember These Definitions

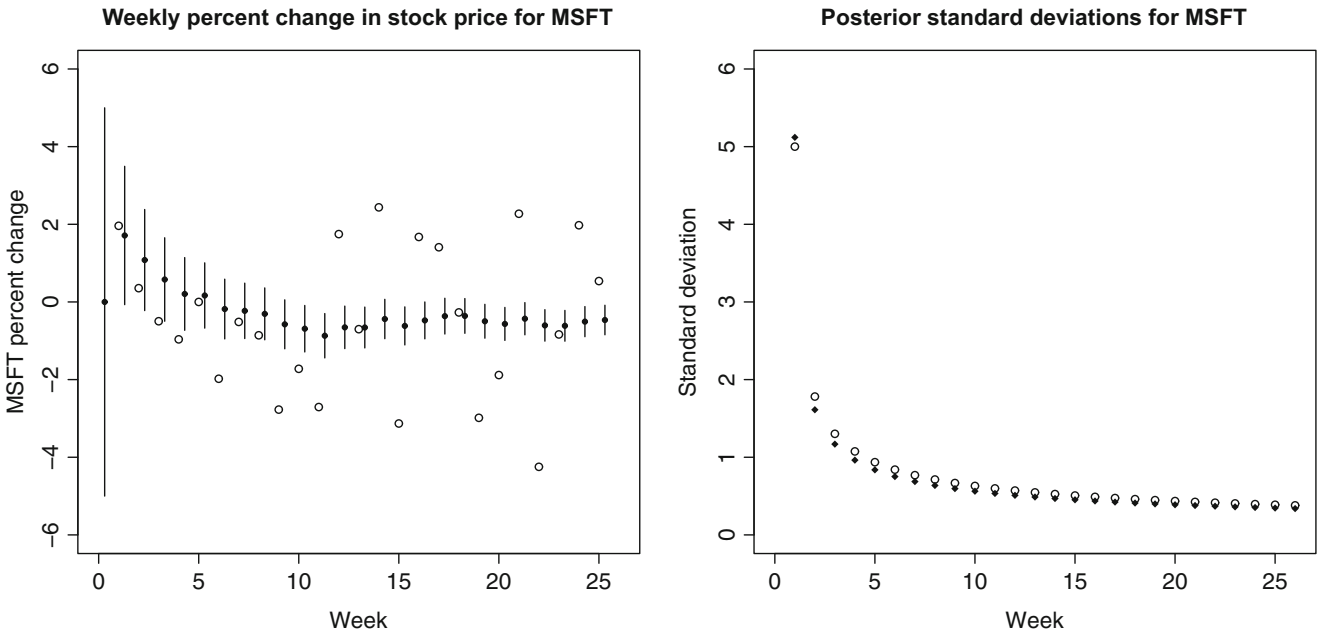

Fig. 9.4 On the *left*, a representation of the posterior on the weekly percent change in price of MSFT stock. The open circles are the observed prices. Distributions are represented by filled circles for the means and bars for the standard deviation (one posterior standard deviation up and down). The first distribution is the prior; then, slightly after each measurement, there's a representation of the posterior for that and all previous measurements. On the *right*, a plot of the posterior standard deviation after the k'th measurement (open circles) together with $1.7/(k+0.11)$ (filled diamonds); the 0.11 is to avoid division by zero for the prior. The close agreement suggests that the k'th standard deviation should be proportional to $1/k$, which simple algebra will confirm

9.4.2 Remember These Terms

9.4.3 Remember These Facts

9.4.4 Use These Procedures

9.4.5 Be Able to

- Write out the likelihood for a set of independent data items produced by models from Chap. 5 (at least Normal, Binomial, Multinomial, Poisson, Beta, Gamma, Exponential).
- Write out the log likelihood for a set of independent data items produced by models from Chap. 5 (at least Normal, Binomial, Multinomial, Poisson, Beta, Gamma, Exponential).
- Find maximum likelihood solutions for parameters of these models from a set of independent data items (in this case, ignore Beta and Gamma; finding maximum likelihood estimates for these can be tricky, and isn't important to us).
- Describe situations where maximum likelihood estimates might not be reliable.
- Describe the difference between maximum likelihood estimation and Bayesian inference.
- Write an expression for the posterior or log-posterior of model parameters given a set of independent data items.
- Compute the MAP estimate for the cases shown in the worked examples.
- Compute on-line estimates of the maximum likelihood estimate of the mean and standard deviation of a normal model.
- Compute on-line estimates of the MAP estimate of the mean and standard deviation in the case of a normal prior and a normal likelihood.

Problems

Maximum Likelihood Methods

9.1 Fitting a Normal Distribution: You are given a dataset of N numbers. Write x_i for the i'th number. You wish to model this dataset with a normal distribution.

(a) Show the maximum likelihood estimate of the mean of this distribution is $\mathsf{mean}\,(\{x\})$.
(b) Show the maximum likelihood estimate of the standard deviation of this distribution is $\mathsf{std}\,(x)$.
(c) Now assume that all of these numbers take the same value—what happens to your estimate of the standard deviation?

9.2 Fitting an Exponential Distribution: You are given a dataset of N non-negative numbers. Write x_i for the i'th number. You wish to model this dataset with an exponential distribution, with probability density function $P(x|\theta) = \theta e^{-\theta x}$. Show the maximum likelihood estimate of θ is

$$\frac{N}{\sum_i x_i}$$

9.3 Fitting a Poisson Distribution: You count the number of times that the annoying "MacSweeper" popup window appears per hour when you surf the web. You wish to model these counts with a Poisson distribution. On day 1, you surf for 4 h, and see counts of $3, 1, 4, 2$ (in hours 1 through 4 respectively). On day 2, you surf for 3 h, and observe counts of $2, 1, 2$. On day 3, you surf for 5 h, and observe counts of $3, 2, 2, 1, 4$. On day 4, you surf for 6 h, but keep only the count for all 6 h, which is 13. You wish to model the intensity in counts per hour.

(a) What is the maximum likelihood estimate of the intensity for each of days 1, 2, and 3 *separately*?
(b) What is the maximum likelihood estimate of the intensity for day 4?
(c) What is the maximum likelihood estimate of the intensity for all days taken together?

9.4 Fitting a Geometric Model: You wish to determine the number of zeros on a roulette wheel without looking at the wheel. You will do so with a geometric model. Recall that when a ball on a roulette wheel falls into a non-zero slot, odd/even bets are paid; when it falls into a zero slot, they are not paid. There are 36 non-zero slots on the wheel.

(a) Assume you observe a total of r odd/even bets being paid before you see a bet not being paid. What is the maximum likelihood estimate of the number of slots on the wheel?
(b) How reliable is this estimate? Why?
(c) You decide to watch the wheel k times to make an estimate. In the first experiment, you see r_1 odd/even bets being paid before you see a bet not being paid; in the second, r_2; and in the third, r_3. What is the maximum likelihood estimate of the number of slots on the wheel?

9.5 Fitting a Binomial Model: You encounter a deck of Martian playing cards. There are 87 cards in the deck. You cannot read Martian, and so the meaning of the cards is mysterious. However, you notice that some cards are blue, and others are yellow.

(a) You shuffle the deck, and draw one card. It is yellow. What is the maximum likelihood estimate of the fraction of blue cards in the deck?
(b) You repeat the previous exercise 10 times, replacing the card you drew each time before shuffling. You see 7 yellow and 3 blue cards in the deck. What is the maximum likelihood estimate of the fraction of blue cards in the deck?

9.6 Fitting a Least Squares Model: We observe a set of N data items. The i'th data item consists of a vector $\mathbf{x}_i$ and a number y_i. We believe that this data is explained by a model where $P(y|\mathbf{x}, \theta)$ is normal, with mean $\mathbf{x}^T\theta$ and (known) standard deviation σ. Here θ is a vector of unknown parameters. At first sight, this model may strike you as being a bit strange, though we'll do a lot with it later. You can visualize this model by noting that y is generated by (a) forming $\mathbf{x}^T\theta$ then (b) adding a zero-mean normal random variable with standard deviation σ.

(a) Show that a maximum likelihood estimate $\hat{\theta}$ of the value of θ is obtained by solving

$$\sum_i (y_i - \mathbf{x}_i^T\hat{\theta})^2 = 0.$$

(b) Stack the y_i into a vector $\mathbf{y}$ and the $\mathbf{x}_i$ into a matrix $\mathcal{X}$ according to

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix} \qquad \mathcal{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \dots \\ \mathbf{x}_N^T \end{bmatrix}.$$

Now show that a maximum likelihood estimate $\hat{\theta}$ of the value of θ is obtained by solving the equation $\mathcal{X}^T\mathcal{X}\hat{\theta} = \mathcal{X}^T\mathbf{y}$.

9.7 Logistic Regression: We observe a set of N data items. The i'th data item consists of a vector $\mathbf{x}_i$ and a discrete y_i, which can take the values 0 or 1. We believe that this data is explained by a model where y_i is a draw from a Bernoulli distribution. The Bernoulli distribution has the property that

$$\log \frac{P(y = 1|\mathbf{x}, \theta)}{P(y = 0|\mathbf{x}, \theta)} = \mathbf{x}^T\theta.$$

This is known as a logistic model. Here θ is a vector of unknown parameters.

(a) Show that

$$P(y = 1|\mathbf{x}, \theta) = \frac{\exp(\mathbf{x}^T\theta)}{1 + \exp(\mathbf{x}^T\theta)}$$

(b) Show that the log-likelihood of a dataset is given by

$$\sum_i \left[y_i \mathbf{x}_i^T \theta - \log\left(1 + \exp(\mathbf{x}_i^T \theta)\right)\right].$$

(c) Now show that a maximum likelihood estimate $\hat{\theta}$ of the value of θ is obtained by solving the equation

$$\sum_i \left[\left(y_i - \frac{\exp(\mathbf{x}_i^T \theta)}{1 + \exp(\mathbf{x}_i^T \theta)} \right) \right] = 0.$$

Likelihood Functions

9.8 You have a dataset of N 1D data items x_i. Each data item is a measurement of the length of an object, in centimeters. You wish to model these items with a normal distribution, with mean μ_c and standard deviation σ_c.

(a) Show that the likelihood, as a function of μ_c and σ_c, is

$$\mathcal{L}_c(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma_c} \prod_i \exp -(x_i - \mu_c)^2 / 2\sigma_c^2.$$

(b) Now assume that you rescale each length by measuring in meters. Your data set now consists of $y_i = 100 x_i$. Show that

$$\begin{aligned} \mathcal{L}_m(\mu_m, \sigma_m) &= \frac{1}{\sqrt{2\pi}\sigma_m} \prod_i \exp -(y_i - \mu_m)^2 / 2\sigma_m^2 \\ &= (1/100)\mathcal{L}_c(\mu_c, \sigma_c) \end{aligned}$$

(c) Use this to argue that the value of the likelihood is not directly meaningful.

Bayesian Methods

9.9 Zeros on a Roulette Wheel: We now wish to make a more sophisticated estimate of the number of zeros on a roulette wheel without looking at the wheel. We will do so with Bayesian inference. Recall that when a ball on a roulette wheel falls into a non-zero slot, odd/even bets are paid; when it falls into a zero slot, they are not paid. There are 36 non-zero slots on the wheel. We assume that the number of zeros is one of $\{0, 1, 2, 3\}$. We assume that these cases have prior probability $\{0.1, 0.2, 0.4, 0.3\}$.

(a) Write n for the event that, in a single spin of the wheel, an odd/even bet will not be paid (equivalently, the ball lands in one of the zeros). Write z for the number of zeros in the wheel. What is $P(n|z)$ for each of the possible values of z (i.e. each of $\{0, 1, 2, 3\}$)?
(b) Under what circumstances is $P(z = 0|\text{observations})$ NOT 0?
(c) You observe 36 independent spins of the same wheel. A zero comes up in 2 of these spins. What is $P(z|\text{observations})$?

9.10 Which random number generator? You have two random number generators. R1 generates numbers that are distributed uniformly and at random in the range −1–1. R2 generates standard normal random variables. A program chooses one of these two random number generators uniformly and at random, then using that generator produces three numbers x_1, x_2 and x_3.

(a) Write $R1$ for the event the program chose R1, etc. When is $P(R1|x_1, x_2, x_3) = 0$?
(b) You observe $x_1 = -0.1$, $x_2 = 0.4$ and $x_3 = -0.9$. What is $P(R1|x_1, x_2, x_3)$?

9.11 Which random number generator, again? You have $r > 1$ random number generators. The i'th random number generator generates numbers that are distributed normally, with zero mean and standard deviation i. A program chooses one of these random number generators uniformly and at random, then using that generator produces N numbers $x_1, \ldots, x_N$. Write R_i for the event the program chose the i'th random number generator, etc. Write an expression for $P(R_i|x_1, \ldots, x_N) = 0$

9.12 Which random number generator, yet again? You have $r > 1$ random number generators. The i'th random number generator generates numbers that are distributed normally, with mean i and standard deviation 2. A program chooses one of these random number generators uniformly and at random, then using that generator produces N numbers $x_1, \ldots, x_N$. Write R_i for the event the program chose the i'th random number generator, etc. Write an expression for $P(R_i|x_1, \ldots, x_N) = 0$

9.13 A Normal Distribution: You are given a dataset of 3 numbers, $-1, 0, 20$. You wish to model this dataset with a normal distribution with unknown mean μ and standard deviation 1. You will make an MAP estimate of μ. The prior on μ is normal, with mean 0 and standard deviation 10.

(a) What is the MAP estimate of μ?
(b) A new datapoint, with value 1, arrives. What is the new MAP estimate of μ?

Bayesian Confidence Intervals

9.14 In Worked example 7.10, we found no reason to reject the idea that mouse weights are normally distributed, using the dataset at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml.

(a) Construct a 75% confidence interval for the weight of a mouse that eats chow, using this data. You should get this interval using reasoning about standard errors—this isn't Bayesian.
(b) Now construct a Bayesian 75% confidence interval for the weight of a mouse that eats chow, using this data. You should assume that the prior on the weight of such a mouse is normal, with mean 32 and standard deviation 10. Recall from Sect. 9.3.2 that a normal prior and a normal likelihood lead to normal posterior.
(c) Compare the intervals constructed above with a Bayesian 75% confidence interval for the weight of a mouse that eats chow, using this data and assuming the prior on the weight of such a mouse is normal, with mean 32 and standard deviation 1. What does this tell you about the importance of the prior?

Programming Exercises

Simulation and Maximum Likelihood

9.15 One interesting way to evaluate maximum likelihood estimation is to use simulations. We will compare maximum likelihood estimates to true parameter values for normal distributions. Write a program that draws s sets of k samples from a normal distribution with mean zero and standard deviation one. Now compute the maximum likelihood estimate of the mean of this distribution from each set of samples. You should see s different values of this estimate, one for each set. How does the variance of the estimate change with k?

9.16 Write a program that draws a sample of a Bernoulli random variable δ with $P(\delta = 1) = 0.5$. If this sample takes the value 1, your program should draw a sample from a normal distribution with mean zero and standard deviation 1. Otherwise, your program should draw a sample from a normal distribution with mean 1 and standard deviation 1. Write x_1 for the resulting sample. We will use x_1 to infer the value of δ_1.

(a) Show that

$$\delta = \begin{cases} 1 \text{ if} x_1 < 0.5 \\ 0 \text{ otherwise} \end{cases}$$

is a maximum likelihood estimate of δ_1.

(b) Now draw 100 sets each of one sample from this program, and infer for each the value of δ_1. How often do you get the right answer?

(c) Show that the true error rate is

$$2\int_{0.5}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{(u-1)^2}{2}\right] du.$$

(Hint: I did a devious change of variables to get the leading 2). Compare the number obtained from your simulation with an estimate of this integral using the error function.

9.17 Logistic regression: We observe a set of N data items. The i'th data item consists of a vector $\mathbf{x}_i$ and a discrete y_i, which can take the values 0 or 1. We believe that this data is explained by a model where y_i is a draw from a Bernoulli distribution. The Bernoulli distribution has the property that

$$\log \frac{P(y=1|\mathbf{x},\theta)}{P(y=0|\mathbf{x},\theta)} = \mathbf{x}^T\theta.$$

This is known as a logistic model. Here θ is a vector of unknown parameters. Inferring θ is often known as logistic regression. We will investigate logistic regression using simulation.

(a) Write a program that accepts ten dimensional vector θ and then (a) generates a sample of 1000 samples $\mathbf{x}_i$, each of which is an IID sample from a normal distribution with mean zero and covariance matrix the identity matrix; and (b) for each $\mathbf{x}_i$, forms y_i which is a sample from a Bernoulli distribution where

$$P(y_i = 1|\mathbf{x}_i,\theta) = \frac{\exp \mathbf{x}_i^T\theta}{1+\exp \mathbf{x}_i^T\theta}.$$

This is a sample dataset. Call this program the dataset maker.

(b) Write a program that accepts a sample dataset and estimates $\hat{\theta}$ (the value of θ that was used to generate the sample dataset) using maximum likelihood. This will require that you use some optimization code, or write your own. Call this program the inference engine.

(c) Choose a θ (a sample from a normal distribution with mean zero and covariance matrix the identity matrix is one way to do this), then create 100 sample datasets. For each, apply the inference engine, and obtain a $\hat{\theta}$. How does the mean of these estimates of θ compare to the true value? What are the eigenvalues of the covariance of these estimates like?

(d) Choose a θ (a sample from a normal distribution with mean zero and covariance matrix the identity matrix is one way to do this), then create 2 sample datasets. For the first, apply the inference engine, and obtain a $\hat{\theta}$. Now use this $\hat{\theta}$ to predict the y_i values of the second, using the logistic regression model. How do your predictions compare to the true values? should the difference be zero? why?

Simulation Based Confidence Intervals

9.18 In the mouse dataset at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml, there are 92 mice that ate a chow diet ("chow" in the relevant column) and where the weight at sacrifice is known (the value of Weight2).

(a) Draw a sample of 30 chow eating mice uniformly at random, and compute the mean weight of these mice. Now use the simulation method of Sect. 9.1.4 to estimate a centered 75% confidence interval for the mean weight of a mouse eating a chow diet, estimated from a sample of 30 mice.

(b) Draw 1000 samples of 30 chow eating mice uniformly at random, and use these samples to estimate a centered 75% confidence interval for the mean weight of a mouse eating a chow diet, estimated from a sample of 30 mice. How does this estimate compare to the previous estimate?

Bayesian Confidence Intervals

9.19 Example 9.5 gives data on swearing by a politician (which I've reproduced below, for your convenience).

No. of swear words	0	1	2	3	4
No. of intervals	5	2	2	1	0

and for the following 20 intervals, you see

No. of swear words	0	1	2	3	4
No. of intervals	9	5	3	2	1

Worked example 9.13 shows how to use a conjugate prior gamma distribution to obtain a gamma posterior. Write a program to identify a centered, 90% bayesian confidence interval for the intensity of the politicians swearing for different values of the prior parameters. How do different choices of these parameters affect your interval?