Contents

CHAPTER 1

Inference Using Models

Assume we have a dataset $\{\mathbf{x}\}$, and a probability model we believe applies to that dataset. For example, we might have a set of N coin flips which we believe to be independent and identically distributed. Of these, k flips came up H. We know that a binomial distribution with $p(H) = p$ is a good model — but what value of p should we use? Your intuition is likely to suggest using k/N, but we'd like a more robust procedure than guessing.

Inference is the process of drawing conclusions from data. We need an inference procedure to obtain the unknown parameter from the data. For some problems, you might just need to know the parameter value that is "best". Such an estimate is known as a **point estimate**. We deal with such problems in this chapter. Notice that this number may not be the right answer; it's just the best estimate of the right answer.

As we shall see, there is more than one possible procedure to apply. Which one we use depends to some extent on the problem. In some cases, we have no particular reason to prefer one value of a parameter to another; in other cases, we might have good reasons to feel some parameter values are more likely than others. For example, if the coin had been borrowed from an acquaintance with an impressive reputation for dishonesty, then you might be willing to believe that p could have almost any value. But if you took a coin at random out of your pocket and flipped that, then you might need a lot of evidence to convince you that p was different from 0.5.

1.1 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 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. 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.

Example: 1.1 *Inferring p from repeated flips — binomial*

We could flip the coin N times, and count the number of heads k. We know that an appropriate probability model for a set of independent coin flips is the binomial model $P(k; N, p)$. But we do not know p, which is the parameter — we need a strategy to extract a value of p from the data.

Example: 1.2 *Inferring p from repeated flips — geometric*

We could flip the coin repeatedly until we see a head. We know that, in this case, the number of flips has the geometric distribution with parameter p. 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 $P_g(N; p)$. But we do not know p, which is the parameter — we need a strategy to extract a value of p from the data.

Example: 1.3 *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 λ? 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 λ from this dataset.

Example: 1.4 *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. We could ask what is the mean and standard deviation of the normal distribution that best represents the data?

We can write that model as $P(\mathcal{D}|\theta)$, where θ are parameters of the probability mode. The model is conditioned on θ, because if we knew θ we could evaluate the model. The expression $P(\mathcal{D}|\theta)$ is known as the **likelihood** of the data, and is often written $\mathcal{L}(\theta)$ (or $\mathcal{L}(\theta; \mathcal{D})$ if you want to remember that data is involved). Notice that this is unlike our models to date. In chapter 13, we assumed that we knew θ, and could then use the model to assign a probability to a data item. Here we *know*

the value of $\mathcal{D}$. The likelihood is a function of θ.

1.1.1 The Maximum Likelihood Principle

We need a "reasonable" procedure to choose a value of θ to report. One — and we stress this is not the only one — is the **maximum likelihood principle**. This says: Choose 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 section **??**). In turn, this means that the likelihood is a product of terms, one for each data item, which we can write as

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

It is traditional to write θ for any set of parameters that are unknown. 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 of that parameter, which we will write $\hat{\theta}$. This estimate is the best we can do — it may not be the "true" value of the parameter.

Worked example 1.1 *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 $P(k; N, \theta)$. We have that

$$\mathcal{L}(\theta) = P(\mathcal{D}|\theta) = P_b(k; N, \theta) = \left(\begin{array}{c} N \\ k \end{array} \right) \theta^k (1-\theta)^{(N-k)}$$

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

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

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

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

This means the maximum likelihood estimate is

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

which is what we guessed would happen, but now we know why that guess "makes sense".

Worked example 1.2 *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

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = 0 = ((1-\theta)^{(N-1)} - (N-1)(1-\theta)^{(n-2)}\theta)$$

So the maximum likelihood estimate is

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

We didn't guess this.

Worked example 1.3 *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 $p_1, \ldots, p_6$.

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,

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

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

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

The logarithm is a monotonic function (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 is very useful, because it allows us to transform a product into a sum. The derivative of a product involves numerous terms; the derivative of a sum is easy to take. We have

$$\log P(\mathcal{D}|\theta) = \log \prod_{i \in \text{dataset}} P(d_i|\theta) = \sum_{i \in \text{dataset}} \log P(d_i|\theta)$$

and in some cases, $\log P(d_i|\theta)$ takes a convenient, easy form. The log-likelihood of a dataset under a model is a function of the unknown parameters, and you will often see it written as

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

Worked example 1.4 *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, λ?

Solution: The likelihood is

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

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 1.5 *Is swearing Poisson?*

A famously sweary politician gives a talk. You listen to the talk, and for each of 30 intervals 1 minute 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	no. of intervals
0	4
1	2
2	2
3	1
4	0

and for the following 20 intervals, you see

no. of swear words	no. of intervals
0	9
1	6
2	3
3	2
4	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 4 to find

$$\begin{aligned}\hat{\lambda}_{10} &= \frac{\text{total number of swearwords}}{\text{number of intervals}} \\ &= \frac{7}{10}\end{aligned}$$

$$\begin{aligned}\hat{\lambda}_{20} &= \frac{\text{total number of swearwords}}{\text{number of intervals}} \\ &= \frac{22}{20}\end{aligned}$$

$$\begin{aligned}\hat{\lambda}_{30} &= \frac{\text{total number of swearwords}}{\text{number of intervals}} \\ &= \frac{29}{30}.\end{aligned}$$

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 1.6 *Normal distributions*

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 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

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

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.

Worked example 1.7 *Normal distributions -II*

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

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

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

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

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.

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.

Another is that, in some cases, you can make online estimates. Assume, rather than seeing N elements of a dataset in one go, you get to see each one once, and you cannot store them. Assume that this dataset is modelled as normal data. Write $\hat{\mu}_k$ for the maximum likelihood estimate of the mean based on data items $1 \ldots k$ (and $\hat{\sigma}_k$ for the maximum likelihood estimate of the standard deviation, etc.). Notice that

$$\hat{\mu}_{k+1} = \frac{(k\hat{\mu}_k) + x_{k+1}}{(k+1)}$$

and that

$$\hat{\sigma}_{k+1} = \sqrt{\frac{(k\hat{\sigma}_k^2) + (x_{k+1} - \hat{\mu}_{k+1})^2}{(k+1)}}$$

This means that you can incorporate new data into your estimate as it arrives without keeping all the data. This process of updating a representation of a dataset as new data arrives is known as **filtering**.

1.1.2 Cautions about Maximum Likelihood

Our examples suggest some difficulties could occur in inference. The first 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. There is a body of mathematics, well outside the scope of this book, that implies that for lots of data that is well described by our model, maximum likelihood will give an answer very close to the "right" answer. This doesn't apply to small datasets. 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. For example, a really important part of natural language processing involves estimating the probability of groups of three words. These groups are usually known as "trigrams". People typically know an awful lot of words (tens to hundreds of thousands, depending on what you mean by a word). This means that there are a tremendous number of trigrams, and you can expect that any real dataset lacks almost all of them, because it isn't big enough. Some are missing because they don't occur in real life, but others are not there simply because they are unusual (eg "Atom Heart Mother" actually occurs in real life, but you may not have seen it all that often). Modern speech recognition systems need to know how probable *every* trigram is. Worse, if a trigram is modelled as having zero probability and actually occurs, the system will make a mistake, so it is important to model all such events as having a very small, but not zero, probability.

In summary, the maximum likelihood estimate is useful, and is consistent with intuition, but small datasets present some worries because there is a real prospect that the best estimate is wrong in a way that presents problems.

1.2 INCORPORATING PRIORS WITH BAYESIAN INFERENCE

Sometimes when we wish to estimate parameters of a model we have prior information. For example, we may have good reason to believe that some parameter is close to some value. 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 $\mathcal{D}$. Extracting information from the posterior is usually called **Bayesian inference**. A natural estimate of θ is the value that maximizes the posterior. This estimate is sometimes known as a **maximum a priori estimate** or **MAP estimate**.

1.2.1 Constructing the Posterior

Bayes' rule tells us that

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

but (as we shall see) it can be hard to work out $P(\mathcal{D})$. For some problems, we might not need to know it.

Worked example 1.8 *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 $0 \leq \theta \leq 1$. It is reasonable then that the prior on p is uniform. We have that $p(\{7 \text{ heads and } 3 \text{ tails}\}|\theta)$ is binomial. Ignore the normalizing constant, and form the joint distribution, which 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:

$$\left(\begin{array}{c} 10 \\ 7 \end{array} \right) \theta^7 (1-\theta)^3$$

which is graphed in figure 1.1. Simply looking at the figure will give some insight into where the probability is in $p(\theta|\{7 \text{ heads and } 3 \text{ tails}\})$. The figure also shows

$$\left(\begin{array}{c} 10 \\ 7 \end{array} \right) \theta^3 (1-\theta)^7$$

which is *proportional* to the posterior for 3 heads and 7 tails. Notice how, 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.

In Example 8, it is interesting to follow how the posterior on p changes as evidence come in, which is easy to do because the posterior is proportional to a binomial distribution. Figure 1.2 shows a set of these posteriors for different sets of evidence.

For other problems, we will need to marginalize out θ, by computing

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

It is usually impossible to do this in closed form, so we would have to use a numerical integral. In some cases, $P(\theta)$ and $P(\mathcal{D}|\theta)$ are **conjugate**, meaning that $P(\theta|\mathcal{D})$ will

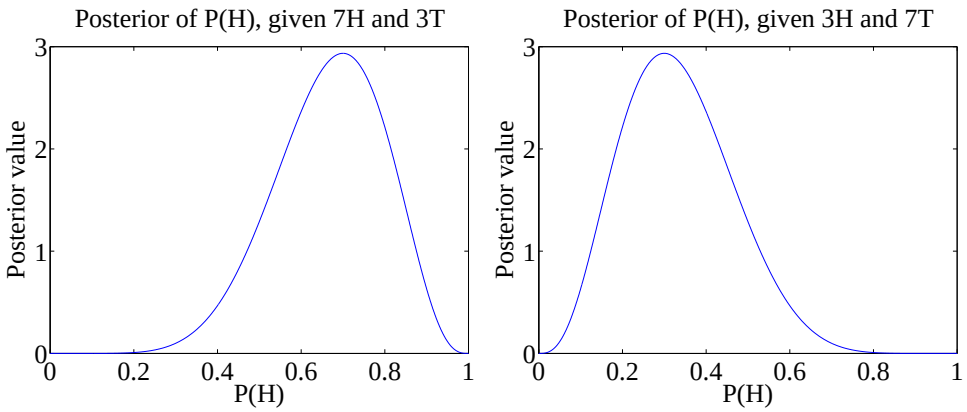

FIGURE 1.1: *The curves show a function* proportional to *the posterior on θ, for the two cases of example 8. Notice that this information is rather richer than the single value we would get from maximum likelihood inference.*

take a familiar form and $P(\mathcal{D})$ follows easily.

Worked example 1.9 *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

$$P(\theta|N, h, \alpha, \beta) \propto \left(\begin{array}{c} N \\ h \end{array} \right) \theta^h (1-\theta)^{(N-h)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{(\alpha-1)} (1-\theta)^{(\beta-1)}.$$

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

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

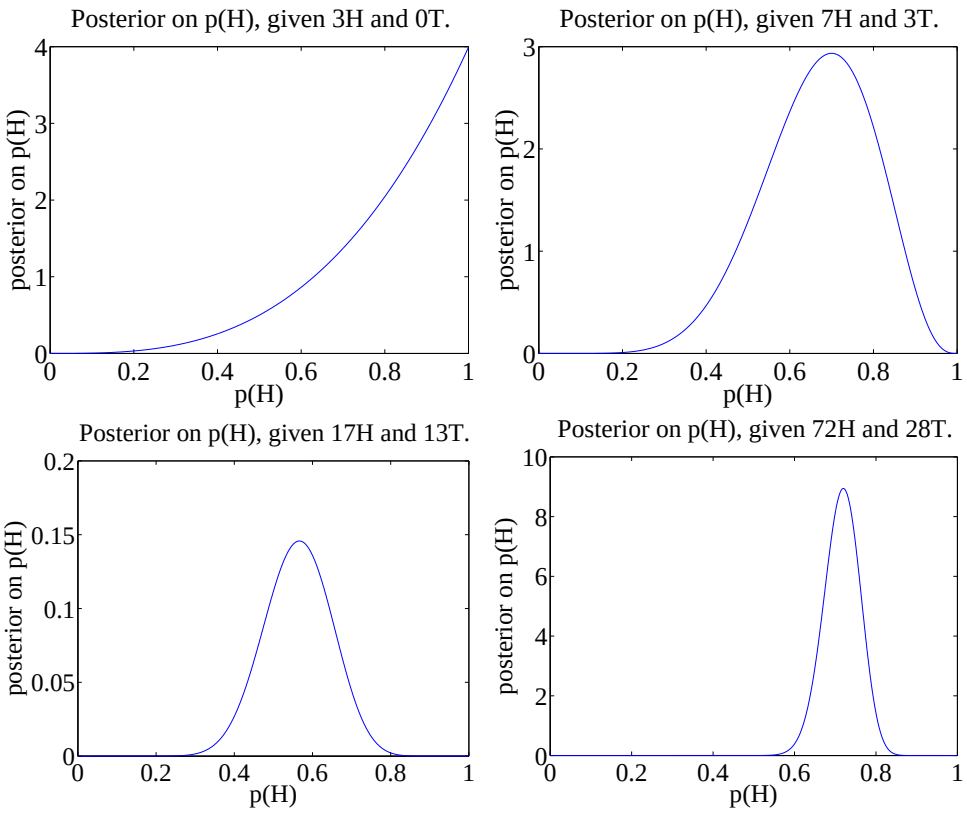

FIGURE 1.2: *The probability that an unknown coin will come up heads when flipped is $p(H)$. For these figures, I simulated coin flips from a coin with $p = 0.75$. I then plotted the posterior for various data. Notice how, as we see more flips, we get more confident about p.*

Worked example 1.10 *More sweary politicians*

Example 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

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

and

$$p(\mathcal{D}|\theta) = \frac{\theta^7 e^{-\theta}}{24}.$$

This means that

$$p(\theta|\mathcal{D}) \propto \theta^{(\alpha-1+7)}e^{-(\beta+1)\theta}.$$

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

$$p(\theta|\mathcal{D}) = \frac{(\beta+1)^{(\alpha+7)}}{\Gamma(\alpha+7)}\theta^{(\alpha-1+7)}e^{-(\beta+1)\theta}$$

1.2.2 The Posterior for Normal Data

There is a very useful construction for the posterior for data where the likelihood is normal. We start with a simple example. Assume we drop a measuring device down a borehole. It is designed to stop falling and catch onto the side of the hole after it has fallen μ_0 meters. On board is a device to measure its depth. This device reports a known constant times the correct depth plus a zero mean normal random variable, which we call "noise". The device reports depth every second.

The first question to ask is what depth do we believe the device is at *before* we receive any measurement? We designed the device to stop at μ_0 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_0 meters plus a zero mean normal random variable. The second term could be caused by error in the braking system, etc. We could estimate the standard deviation of the second term (which we write σ_0) 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_0 and standard deviation σ_0.

Notice that this model probably isn't exactly right — for example, there must be some probability in the model that the object falls beyond the bottom of the hole, which it can't do — but it captures some important properties of our system. The device should stop at or close to μ_0 meters most of the time, and it's unlikely to be too far away.

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 — our original guess about the device's location might be better than the measurement. Write x_1 for the measurement. Notice that the scale of the measurement may not be the same as the scale of the depth, so the mean of the measurement is $c_1\theta$, where c_1 is a change of scale (for example, from inches to meters). We have that $p(x_1|\theta)$ is normal with mean $c_1\theta$ and standard deviation σ_{n1}. We would like to know $p(\theta|x_1)$.

We have that

$$\begin{aligned}\log p(\theta, x_1) &= \log p(x_1|\theta) + \log p(\theta) \\ &= -\frac{(x_1 - c_1\theta)^2}{2\sigma_{n1}^2} - \frac{(\theta - \mu_0)^2}{2\sigma_0^2} \\ &+ \text{terms not depending on } \theta \text{ or } x.\end{aligned}$$

We have two estimates of the position, θ, and we wish to come up with a representation of what we know about θ. One is x_1, which is a *measurement* — we know its value. The expected value of x_1 is $c_1\theta$, so we could infer θ from x_1. But we have another estimate of the position, which is μ_0. The posterior, $p(\theta|x_1)$, is a probability distribution on the variable θ; it depends on the known values x_1, μ_0,

σ_0 and σ_{n1}. We need to determine its form. We can do so by some rearrangement of the expression for $\log p(\theta, x_1)$.

Notice first that this expression 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. This yields a fact of crucial importance.

Useful Fact: 1.1 *Normal distributions are conjugate*

A normal prior and a normal likelihood yield a normal posterior.

Write μ_1 for the mean of this distribution, and σ_{n1} for its standard deviation. The log of the distribution must be

$$-\frac{(\theta-\mu_1)^2}{2\sigma_1^2} + \text{terms not depending on } \theta.$$

The terms not depending on θ are not interesting, because if we know σ_1 those terms must add up to

$$\log\left(\frac{1}{\sqrt{2\pi}\sigma_1}\right)$$

so that the probability density function sums to one. Our goal is to rearrange terms into the form above. Notice that

$$-\frac{(\theta-\mu_1)^2}{2\sigma_p^2} = -\theta^2\left(\frac{1}{2\sigma_1^2}\right) + 2\theta\frac{\mu_1}{2\sigma_p^2} + \text{term not depending on } \theta$$

We have

$$\begin{aligned}
\log p(\theta|x_1) &= -\frac{(c_1\theta - x_1)^2}{2\sigma_{n1}^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2} + \text{terms not depending on } \theta \\
&= -\theta^2\left(\frac{1}{2\left(\frac{\sigma_{n1}^2\sigma_0^2}{\sigma_{n1}^2+c_1^2\sigma_0^2}\right)}\right) + 2\theta\left(c_1\frac{x_1}{2\sigma_{n1}^2} + \frac{\mu_0}{2\sigma_0^2}\right) \\
&+ \quad \text{terms not depending on } \theta
\end{aligned}$$

which means that

$$\sigma_1^2 = \frac{\sigma_{n1}^2\sigma_0^2}{\sigma_{n1}^2 + c_1^2\sigma_0^2}$$

and

$$\begin{aligned}
\mu_1 &= 2\left(c_1\frac{x_1}{2\sigma_{n1}^2} + \frac{\mu_0}{2\sigma_0^2}\right)\frac{\sigma_{n1}^2\sigma_0^2}{\sigma_{n1}^2 + c_1^2\sigma_0^2} \\
&= \left(\frac{c_1x_1\sigma_0^2 + \mu_0\sigma_{n1}^2}{\sigma_{n1}^2\sigma_0^2}\right)\frac{\sigma_{n1}^2\sigma_0^2}{\sigma_{n1}^2 + c_1^2\sigma_0^2} \\
&= \frac{c_1x_1\sigma_0^2 + \mu_0\sigma_{n1}^2}{\sigma_{n1}^2 + c_1^2\sigma_0^2}.
\end{aligned}$$

These equations is that they "make sense". Imagine that σ_0 is very small, and σ_{n1} is very big; then our new expected value of θ — which is μ_1 — is about μ_0. Equivalently, because our prior was very accurate, and the measurement was unreliable, our expected value is about the prior value. Similarly, if the measurement is reliable (i.e. σ_{n1} is small) and the prior has high variance (i.e. σ_0 is large), then our expected value of θ is about x_1/c_1 — i.e. the measurement, rescaled. I have put these equations, in a more general form, in a box below.

Useful Fact: 1.2 *Normal posteriors*

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. The likelihood of this data item is normal with mean $c\theta$ and standard deviation σ_m, where c and σ_m are known. Then the posterior, $p(\theta|x, c, \sigma_m, \mu_\pi, \sigma_\pi)$, is normal, with mean

$$\frac{cx\sigma_\pi^2 + \mu_\pi\sigma_m^2}{\sigma_m^2 + c^2\sigma_\pi^2}$$

and standard deviation

$$\sqrt{\frac{\sigma_m^2\sigma_\pi^2}{\sigma_m^2 + c^2\sigma_\pi^2}}.$$

Assume a second measurement, x_2 arrives. We know that $p(x_2|\theta, c_2, \sigma_{n2})$ is normal with mean $c_2\theta$ and standard deviation σ_{n2}. In the example, we have a new measurement of depth — perhaps in a new, known, scale — with new noise (which might have larger, or smaller, standard deviation than the old noise) added. Then we can use $p(\theta|x_1, c_1, \sigma_{n1})$ as a prior to get a posterior $p(\theta|x_1, x_2, c_1, c_2, \sigma_{n1}, \sigma_{n2})$. Each is normal, by useful fact 1. Not only that, but we can easily obtain the expressions for the mean μ_2 and the standard deviation σ_2 *recursively* as functions of μ_1 and σ_1.

Applying useful fact 2, we have

$$\mu_2 = \frac{c_2x_2\sigma_1^2 + \mu_1\sigma_{n2}^2}{\sigma_{n2}^2 + c_2^2\sigma_1^2}$$

and

$$\sigma_2^2 = \frac{\sigma_{n2}^2\sigma_1^2}{\sigma_{n2}^2 + c_2^2\sigma_1^2}.$$

But what works for 2 and 1 will work for $k+1$ and k. We know the posterior after k measurements will be normal, with mean μ_k and standard deviation σ_k. The $k+1$'th measurement x_{k+1} arrives, and we have $p(x_{k+1}|\theta, c_{k+1}, \sigma_{n(k+1)})$ is normal. Then the posterior is normal, and we can write the mean μ_{k+1} and the standard

deviation σ_{k+1} recursively as functions of μ_k and σ_k. This yields

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

and

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

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.

1.2.3 MAP Inference

Look at example 13, 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 $P(\mathcal{D}, \theta)$, often called the **joint** distribution.

Worked example 1.11 *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 1.12 *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

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

and the MAP estimate is

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

(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.

Worked example 1.13 *Normal data*

Assume you see N datapoints x_i which are modelled by a normal distribution with unknown mean θ and with known standard deviation σ. You model the prior on θ using a normal distribution with mean μ_0 and standard deviation σ_0. What is the MAP estimate of the mean?

Solution: Recall that the maximum value of a normal distribution occurs at its mean. Now problem is covered by useful fact 2, but in this case we have $c_i = 1$ for each data point, and $\sigma_i = \sigma$. We can write

$$\mu_N = \frac{x_N\sigma_{N-1}^2 + \mu_{N-1}\sigma^2}{\sigma^2 + \sigma_{N-1}^2}$$

and

$$\sigma_N^2 = \frac{\sigma^2\sigma_{N-1}^2}{\sigma^2 + \sigma_{N-1}^2}.$$

and evaluate the recursion down to μ_0, σ_0.

1.2.4 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 inference 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 succesful 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.