Choosing priors
Class 16, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Learn that the choice of prior affects the posterior.
2. See that too rigid a prior can make it difficult to learn from the data.
3. See that more data lessons the dependence of the posterior on the prior.
4. Be able to make a reasonable choice of prior, based on prior understanding of the system under consideration.

2 Introduction

Up to now we have always been handed a prior pdf. In this case, statistical inference from data is essentially an application of Bayes' theorem. When the prior is known there is no controversy on how to proceed. The art of statistics starts when the prior is not known with certainty. There are two main schools on how to proceed in this case: Bayesian and frequentist. For now we are following the Bayesian approach. Starting next week we will learn the frequentist approach.

Recall that given data D and a hypothesis H we used Bayes' theorem to write

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}.$$

Bayesian: Bayesians make inferences using the posterior $P(H|D)$, and therefore always need a prior $P(H)$. If a prior is not known with certainty the Bayesian must try to make a reasonable choice. There are many ways to do this and reasonable people might make different choices. In general it is good practice to justify your choices and to explore a range of priors to see if they all point to the same conclusion.

Frequentist: Very briefly, frequentists do not try to create a prior. Instead, they make inferences using the likelihood $P(D|H)$.

We will compare the two approaches in detail once we have more experience with each. For now we simply list two benefits of the Bayesian approach.

1. The posterior probability $P(H|D)$ for the hypothesis given the evidence is usually exactly what we'd like to know. The Bayesian can say something like 'the parameter of interest has probability 0.95 of being between 0.49 and 0.51.'
2. The assumptions that go into choosing the prior can be clearly spelled out.

More good data: It is always the case that more good data allows for stronger conclusions and lessens the influence of the prior. The emphasis should be as much on good data (quality) as on more data (quantity).

3 Example: Dice

Suppose we have a drawer full of dice, each of which has either 4, 6, 8, 12, or 20 sides. This time, we do not know how many of each type are in the drawer. A die is picked at random from the drawer and rolled 5 times. The results in order are 4, 2, 4, 7, and 5.

3.1 Uniform prior

Suppose we have no idea what the distribution of dice in the drawer might be. In this case it's reasonable to use a flat prior. Here is the update table for the posterior probabilities that result from updating after each roll. In order to fit all the columns, we leave out the unnormalized posteriors.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	1/5	1/4	0.370	1/4	0.542	1/4	0.682	0	0.000	0	0.000
H_6	1/5	1/6	0.247	1/6	0.241	1/6	0.202	0	0.000	1/6	0.000
H_8	1/5	1/8	0.185	1/8	0.135	1/8	0.085	1/8	0.818	1/8	0.876
H_{12}	1/5	1/12	0.123	1/12	0.060	1/12	0.025	1/12	0.161	1/12	0.115
H_{20}	1/5	1/20	0.074	1/20	0.022	1/20	0.005	1/20	0.021	1/20	0.009

This should look familiar. Given the data the final posterior is heavily weighted towards hypthesis H_8 that the 8-sided die was picked.

3.2 Other priors

To see how much the above posterior depended on our choice of prior, let's try some other priors. Suppose we have reason to believe that there are ten times as many 20-sided dice in the drawer as there are each of the other types. The table becomes:

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.071	1/4	0.222	1/4	0.453	1/4	0.650	0	0.000	0	0.000
H_6	0.071	1/6	0.148	1/6	0.202	1/6	0.193	0	0.000	1/6	0.000
H_8	0.071	1/8	0.111	1/8	0.113	1/8	0.081	1/8	0.688	1/8	0.810
H_{12}	0.071	1/12	0.074	1/12	0.050	1/12	0.024	1/12	0.136	1/12	0.107
H_{20}	0.714	1/20	0.444	1/20	0.181	1/20	0.052	1/20	0.176	1/20	0.083

Even here the final posterior is heavily weighted to the hypothesis H_8.

What if the 20-sided die is 100 times more likely than each of the others?

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0.0096	1/4	0.044	1/4	0.172	1/4	0.443	0	0.000	0	0.000
H_6	0.0096	1/6	0.030	1/6	0.077	1/6	0.131	0	0.000	1/6	0.000
H_8	0.0096	1/8	0.022	1/8	0.043	1/8	0.055	1/8	0.266	1/8	0.464
H_{12}	0.0096	1/12	0.015	1/12	0.019	1/12	0.016	1/12	0.053	1/12	0.061
H_{20}	0.9615	1/20	0.889	1/20	0.689	1/20	0.354	1/20	0.681	1/20	0.475

With such a strong prior belief in the 20-sided die, the final posterior gives a lot of weight to the theory that the data arose from a 20-sided die, even though it extremely unlikely the

20-sided die would produce a maximum of 7 in 5 roles. The posterior now gives roughly even odds that an 8-sided die versus a 20-sided die was picked.

3.3 Rigid priors

Mild cognitive dissonance. Too rigid a prior belief can overwhelm any amount of data. Suppose I've got it in my head that the die has to be 20-sided. So I set my prior to $P(H_{20}) = 1$ with the other 4 hypotheses having probability 0. Look what happens in the update table.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	post_4	lik_5	post_5
H_4	0	1/4	0	1/4	0	1/4	0	0	0	0	0
H_6	0	1/6	0	1/6	0	1/6	0	0	0	1/6	0
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	1/8	0
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	1/12	0
H_{20}	1	1/20	1	1/20	1	1/20	1	1/20	1	1/20	1

No matter what the data, a hypothesis with prior probability 0 will have posterior probability 0. In this case I'll never get away from the hypothesis H_{20}, although I might experience some mild cognitive dissonance.

Severe cognitive dissonance. Rigid priors can also lead to absurdities. Suppose I now have it in my head that the die must be 4-sided. So I set $P(H_4) = 1$ and the other prior probabilities to 0. With the given data on the fourth roll I reach an impasse. A roll of 7 can't possibly come from a 4-sided die. Yet this is the only hypothesis I'll allow. My unnormalized posterior is a column of all zeros which cannot be normalized.

hyp.	prior	lik_1	post_1	lik_2	post_2	lik_3	post_3	lik_4	unnorm. post_4	post_4
H_4	1	1/4	1	1/4	1	1/4	1	0	0	???
H_6	0	1/6	0	1/6	0	1/6	0	0	0	???
H_8	0	1/8	0	1/8	0	1/8	0	1/8	0	???
H_{12}	0	1/12	0	1/12	0	1/12	0	1/12	0	???
H_{20}	0	1/20	0	1/20	0	1/20	0	1/20	0	???

I must adjust my belief about what is possible or, more likely, I'll suspect you of accidently or deliberately messing up the data.

4 Example: Malaria

Here is a real example adapted from *Statistics, A Bayesian Perspective* by Donald Berry:

By the 1950's scientists had begun to formulate the hypothesis that carriers of the sickle-cell gene were more resistant to malaria than noncarriers. There was a fair amount of circumstantial evidence for this hypothesis. It also helped explain the persistance of an otherwise deleterious gene in the population. In one experiment scientists injected 30 African volunteers with malaria. Fifteen of the volunteers carried one copy of the sickle-cell gene and the other 15 were noncarriers. Fourteen out of 15 noncarriers developed malaria while only 2

out of 15 carriers did. Does this small sample support the hypothesis that the sickle-cell gene protects against malaria?

Let S represent a carrier of the sickle-cell gene and N represent a non-carrier. Let $D+$ indicate developing malaria and $D-$ indicate not developing malaria. The data can be put in a table.

	$D+$	$D-$	
S	2	13	15
N	14	1	15
	16	14	30

Before analysing the data we should say a few words about the experiment and experimental design. First, it is clearly unethical: to gain some information they infected 16 people with malaria. We also need to worry about bias. How did they choose the test subjects. Is it possible the noncarriers were weaker and thus more susceptible to malaria than the carriers? Berry points out that it is reasonable to assume that an injection is similar to a mosquito bite, but it is not guaranteed. This last point means that if the experiment shows a relation between sickle-cell and protection against injected malaria, we need to consider the hypothesis that the protection from mosquito transmitted malaria is weaker or non-existent. Finally, we will frame our hypothesis as 'sickle-cell protects against malaria', but really all we can hope to say from a study like this is that 'sickle-cell is correlated with protection against malaria'.

Model. For our model let θ_S be the probability that an injected carrier S develops malaria and likewise let θ_N be the probabilitiy that an injected noncarrier N develops malaria. We assume independence between all the experimental subjects. With this model, the likelihood is a function of both θ_S and θ_N:

$$P(\text{data}|\theta_S, \theta_N) = c\,\theta_S^2(1-\theta_S)^{13}\theta_N^{14}(1-\theta_N).$$

As usual we leave the constant factor c as a letter. (It is a product of two binomial coefficients: $c = \binom{15}{2}\binom{15}{14}$.)

Hypotheses. Each hypothesis consists of a pair (θ_N, θ_S). To keep things simple we will only consider a finite number of values for these probabilities. We could easily consider many more values or even a continuous range of hypotheses. Assume θ_S and θ_N are each one of 0, 0.2, 0.4, 0.6, 0.8, 1. This leads to two-dimensional tables.

First is a table of hypotheses. The color coding indicates the following:
1. Light orange squares along the diagonal are where $\theta_S = \theta_N$, i.e. sickle-cell makes no difference one way or the other.
2. Pink and red squares above the diagonal are where $\theta_N > \theta_S$, i.e. sickle-cell provides some protection against malaria.
3. In the red squares $\theta_N - \theta_S \geq 0.6$, i.e. sickle-cell provides a lot of protection.
4. White squares below diagonal are where $\theta_S > \theta_N$, i.e. sickle-cell actually increases the probability of developing malaria.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	(0,1)	(.2,1)	(.4,1)	(.6,1)	(.8,1)	(1,1)
0.8	(0,.8)	(.2,.8)	(.4,.8)	(.6,.8)	(.8,.8)	(1,.8)
0.6	(0,.6)	(.2,.6)	(.4,.6)	(.6,.6)	(.8,.6)	(1,.6)
0.4	(0,.4)	(.2,.4)	(.4,.4)	(.6,.4)	(.8,.4)	(1,.4)
0.2	(0,.2)	(.2,.2)	(.4,.2)	(.6,.2)	(.8,.2)	(1,.2)
0	(0,0)	(.2,0)	(.4,0)	(.6,0)	(.8,0)	(1,0)

Hypotheses on level of protection due to S:
red = strong; pink = some; orange = none; white = negative.

Next is the table of likelihoods. (Actually we've taken advantage of our indifference to scale and scaled all the likelihoods by $100000/c$ to make the table more presentable.) Notice that, to the precision of the table, many of the likelihoods are 0. The color coding is the same as in the hypothesis table. We've highlighted the biggest likelihoods with a blue border.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	1.93428	0.18381	0.00213	0.00000	0.00000
0.6	0.00000	0.06893	0.00655	0.00008	0.00000	0.00000
0.4	0.00000	0.00035	0.00003	0.00000	0.00000	0.00000
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

Likelihoods $p(\text{data}|\theta_S, \theta_N)$ scaled by $100000/c$

4.1 Flat prior

Suppose we have no opinion whatsoever on whether and to what degree sickle-cell protects against malaria. In this case it is reasonable to use a flat prior. Since there are 36 hypotheses each one gets a prior probability of 1/36. This is given in the table below. Remember each square in the table represents one hypothesis. Because it is a probability table we include the marginal pmf.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.8	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.6	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.4	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0.2	1/36	1/36	1/36	1/36	1/36	1/36	1/6
0	1/36	1/36	1/36	1/36	1/36	1/36	1/6
$p(\theta_S)$	1/6	1/6	1/6	1/6	1/6	1/6	1

Flat prior $p(\theta_S, \theta_N)$: every hypothesis (square) has equal probability

To compute the posterior we simply multiply the likelihood table by the prior table and

normalize. Normalization means making sure the entire table sums to 1.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88075	0.08370	0.00097	0.00000	0.00000	0.96542
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00002	0.00000	0.00000	0.00000	0.00018
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91230	0.08670	0.00100	0.00000	0.00000	1.00000

Posterior to flat prior: $p(\theta_S, \theta_N \mid \text{data})$

To decide whether S confers protection against malaria, we compute the posterior probabilities of 'some protection' and of 'strong protection'. These are computed by summing the corresponding squares in the posterior table.

Some protection: $P(\theta_N > \theta_S) = \text{sum of pink and red} = .99995$

Strong protection: $P(\theta_N - \theta_S > .6) = \text{sum of red} = .88075$

Working from the flat prior, it is effectively certain that sickle-cell provides some protection and very probable that it provides strong protection.

4.2 Informed prior

The experiment was not run without prior information. There was a lot of circumstantial evidence that the sickle-cell gene offered some protection against malaria. For example it was reported that a greater percentage of carriers survived to adulthood.

Here's one way to build an informed prior. We'll reserve a reasonable amount of probability for the hypotheses that S gives no protection. Let's say 24% split evenly among the 6 (orange) cells where $\theta_N = \theta_S$. We know we shouldn't set any prior probabilities to 0, so let's spread 6% of the probability evenly among the 15 white cells below the diagonal. That leaves 70% of the probability for the 15 pink and red squares above the diagonal.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N)$
1	0.04667	0.04667	0.04667	0.04667	0.04667	0.04000	0.27333
0.8	0.04667	0.04667	0.04667	0.04667	0.04000	0.00400	0.23067
0.6	0.04667	0.04667	0.04667	0.04000	0.00400	0.00400	0.18800
0.4	0.04667	0.04667	0.04000	0.00400	0.00400	0.00400	0.14533
0.2	0.04667	0.04000	0.00400	0.00400	0.00400	0.00400	0.10267
0	0.04000	0.00400	0.00400	0.00400	0.00400	0.00400	0.06000
$p(\theta_S)$	0.27333	0.23067	0.18800	0.14533	0.10267	0.06000	1.0

Informed prior $p(\theta_S, \theta_N)$: makes use of prior information that sickle-cell is protective.

We then compute the posterior pmf.

$\theta_N \backslash \theta_S$	0	0.2	0.4	0.6	0.8	1	$p(\theta_N \mid \text{data})$
1	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0.8	0.00000	0.88076	0.08370	0.00097	0.00000	0.00000	0.96543
0.6	0.00000	0.03139	0.00298	0.00003	0.00000	0.00000	0.03440
0.4	0.00000	0.00016	0.00001	0.00000	0.00000	0.00000	0.00017
0.2	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
0	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$p(\theta_S \mid \text{data})$	0.00000	0.91231	0.08669	0.00100	0.00000	0.00000	1.00000

Posterior to informed prior: $p(\theta_S, \theta_N|\text{data})$

We again compute the posterior probabilities of 'some protection' and 'strong protection'.

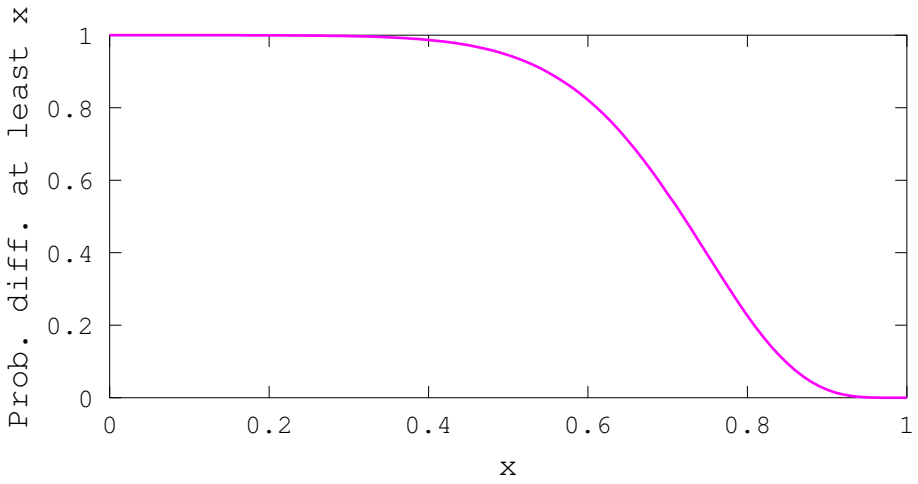
Some protection: $P(\theta_N > \theta_S) =$ sum of pink and red $= .99996$

Strong protection: $P(\theta_N - \theta_S > .6) =$ sum of red $= .88076$

Note that the informed posterior is nearly identical to the flat posterior.

4.3 PDALX

The following plot is based on the flat prior. For each x, it gives the probability that $\theta_N - \theta_S \geq x$. To make it smooth we used many more hypotheses.



Probability the difference $\theta_N - \theta_S$ is at least x (PDALX).

Notice that it is virtually certain that the difference is at least .4.