CHAPTER 7

PRIOR INFORMATION

7.1. Exact prior information

7.1.1. *Prior information*

When there is exact prior information there is no discussion about Bayesian methods and it can be integrated using the rules of probability. The following example is based on an example prepared for Fisher, who was a notorious anti-Bayesian, but he never objected the use of prior probability when clearly established.

There is a type of laboratoy mouse whose skin colour is controlled by a single gene with two alleles 'A' and 'a' so that when the mouse has two copies of the recessive allele (aa) its skin is brown, and it is black in the other cases (AA and aa). We cross two heterozygous and we have a descent that is black coloured. We want to know whether this black mouse is homozygous (AA) or heterozygous (Aa or aA). In order to know this, we cross the mouse with a brown mouse (aa) and examine the offspring (Figure 4.1).

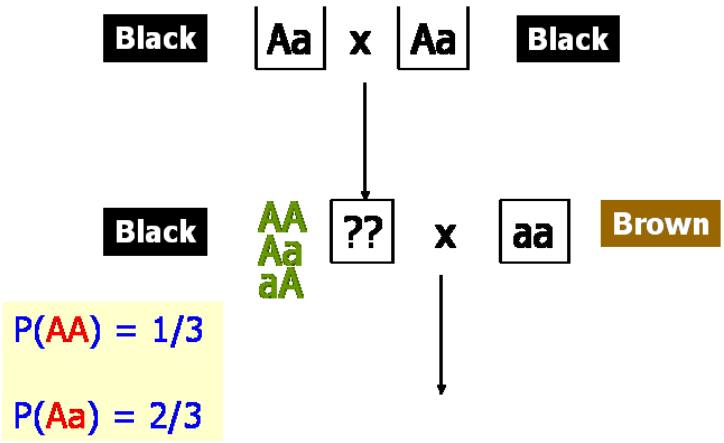

Figure 4.1. Experiment to determine wheter a parent is homocygous (AA) ot heterozygous (Aa or aA)

If we get a brown mouse in the offspring, we will know that the mouse is heterozygous, but if we only get black offspring we still will doubt about wheter it is homo- or heterozygous. If we get many black offspring, it will be unlikely that the mouse is heterozygous, but *before making the experiment* we have some probabilities of obtaining black or brown offspring. We know that the mouse to be tested cannot be 'aa' because otherwise it would be brown, thus it received both alleles 'A' from its mother and father, or an allele 'A' from the father and an allele 'a' from the mother to become 'Aa', or the opposite, to become 'aA'. We have three possibilites, thus the probability of being 'AA' is 1/3 and the probabilities of being heterocygous ('Aa' or 'aA', both are genetically identical ([1])) is 2/3. This is what we expect before having any data from the experiment. Notice that these expectations are not merely 'beliefs', but quantified probabilities. Also notice that they come from our knowledge of the Mendel laws and from the knowledge that our mouse is the son of two heterozygous.

7.1.2. *Posterior probabilities with exact prior information*

[1] We will no make any difference from Aa and aA in the rest of the chapter, thus 'Aa' will mean both 'Aa' and 'aA' from now.

Now, the experiment is made and we obtain three offspring, all black (figure 4.2). They received for sure an allele ‘a’ from the mother and an allele ‘A’ from our mouse, but our mouse still can be homozygous (AA) or heterozygous (Aa). Which is the probability of being each type?

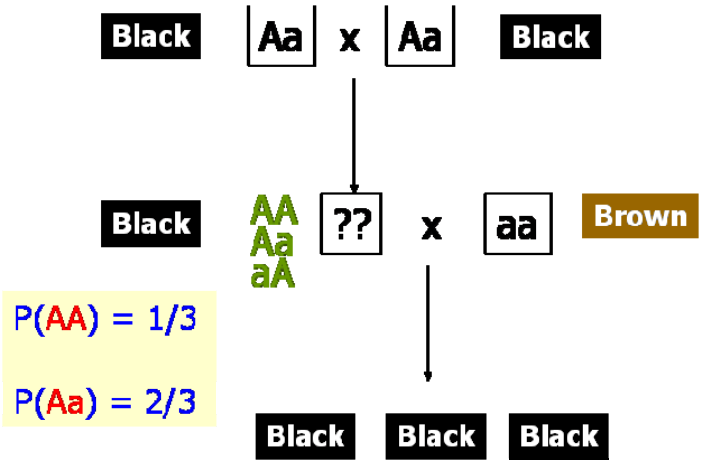

Figure 4.2. Experiment to determine wheter a parent is homocygous (AA) ot heterozygous (Aa or aA)

To know this we will apply Bayes Theorem. The probability of being homozygous (AA) given that we have obtained three offspring black is

$$P(AA \mid \mathbf{y} = 3 \text{ black}) = \frac{P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA)}{P(\mathbf{y} = 3 \text{ black})}$$

We know that *if it is true that our mouse is AA*, the probability of obtaining a black offspring is 1, since the offspring will always have an allele ‘A’. Thus,

$$P(\mathbf{y} = 3 \text{ black} \mid AA) = 1$$

We also know that the prior probability of being AA is 1/3, thus

$$P(AA) = 0.33$$

Finally, the probability of the sample is the sum of the probabilities of two excluding events: having a parent homozygous (AA) or having a parent heterozygous (Aa) (see footnote 2).

$$P(\mathbf{y} = 3 \text{ black}) = P(\mathbf{y} = 3 \text{ black } \& \ AA) + P(\mathbf{y} = 3 \text{ black } \& \ Aa) =$$

$$= P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA) + P(\mathbf{y} = 3 \text{ black} \mid Aa) \cdot P(Aa)$$

to calculate it we need the prior probability of being heterozygous, that we know it is

$$P(Aa) = \frac{2}{3}$$

and the probability of obtaining our sample *if it is true that our mouse is Aa.* If our mouse would be Aa, the only way of obtaining a black offspring is that this offspring get his allele A from him, thus the probability of obtaining one black offspring will be ½. The probability of obtraining three black offspring will be ½ x ½ x ½ , thus

$$P(\mathbf{y} = 3 \text{ black} \mid Aa) = \left(\frac{1}{2}\right)^3$$

Now we can calculate the probability of our sample:

$$P(\mathbf{y} = 3 \text{ black}) = P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA) + P(\mathbf{y} = 3 \text{ black} \mid Aa) \cdot P(Aa) =$$

$$= \quad 1 \quad \cdot \quad \frac{1}{3} \quad + \quad \left(\frac{1}{2}\right)^3 \quad \cdot \quad \frac{2}{3} \quad = 0.42$$

Then, applying Bayes theorem

$$P(AA \mid \mathbf{y} = 3 \text{ black}) = \frac{P(\mathbf{y} = 3 \text{ black} \mid AA) \cdot P(AA)}{P(\mathbf{y} = 3 \text{ black})} = \frac{1 \times 0.33}{0.42} = 0.80$$

The probability of being heterozygous can be calculated again using Bayes theorem, or simplily as

$$P(Aa \mid \mathbf{y} = 3 \text{ black}) = 1 - P(AA \mid \mathbf{y} = 3 \text{ black}) = 1 - 0.80 = 0.20$$

Thus, we had a prior probability, before obtaining any data, and a probability after obtaining three black offspring

$$\text{prior } P(AA) = 0.33 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.80$$

$$\text{prior } P(Aa) = 0.67 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.20$$

before the experiment was performed it was more probable that our mouse was heterozygous (Aa), but after the experiment it is more probable that it is homozygous (AA).

Notice that the sum of both probabilities is 1

P(AA | **y**) + P(Aa | **y**) = 1.00

thus the posterior probabilitites give a relative measure of uncertainty (80% and 20% respectively). However, the sum of the likelihoods is not 1 because they come from different events

P(**y** |AA) + P(**y** |Aa) = 1.125

thus the likelihoods do not provide a *measure* of uncertainty.

7.1.3. *Influence of prior information in posterior probabilitites*

If instead of using exact prior information we had used flat priors, repeating the calculus, we will obtain

$$\text{prior } P(AA) = 0.50 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.89$$
$$\text{prior } P(Aa) = 0.50 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.11$$

we can see that flat prior information had an influence in the final result. When having exact prior information it is better to use it.

If we have exact prior information and we have a large amount of information, for example P(AA) = 0.002, computing the probabilities again we obtain

$$\text{prior } P(AA) = 0.002 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.02$$
$$\text{prior } P(Aa) = 0.998 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.98$$

thus despite of having evidence from the data in favour of AA, we decide that the mouse is Aa because prior information dominates and the posterior distribution favours Aa. This has been a frequent criticism to Bayesian inference, but one wonders why an experiment should be performed when the previous evidence is so strong in favour of Aa.

What could have happened if instead three black offspring we had obtained seven black offspring?

Repeating the calculus for **y** = 7 black, we obtain

$$\text{prior } P(AA) = 0.33 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.99$$
$$\text{prior } P(Aa) = 0.67 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.01$$

If flat priors were used, we obtain

$$\text{prior } P(AA) = 0.50 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.99$$
$$\text{prior } P(Aa) = 0.50 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.01$$

in this case the evidence provided by the data dominates over the prior information. However, if prior information is very large

$$\text{prior } P(AA) = 0.002 \qquad \text{posterior } P(AA|\mathbf{y}) = 0.33$$
$$\text{prior } P(Aa) = 0.998 \qquad \text{posterior } P(Aa|\mathbf{y}) = 0.67$$

thus even having more data, prior information dominates the final result when it is very large, which should not be normally the case. In general, prior information loses importance whith larger samples. For example, if we have n uncorrelated data,

$$f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta) f(\theta) = f(y_1, y_2, \ldots, y_n|\theta) f(\theta) = f(\mathbf{y}_1|\theta) f(\mathbf{y}_2|\theta) \ldots f(\mathbf{y}_n|\theta) f(\theta)$$

taking logarithms

$$\log f(\theta|\mathbf{y}) \propto \log f(\mathbf{y}_1|\theta) + \log f(\mathbf{y}_2|\theta) + \ldots + \log f(\mathbf{y}_n|\theta) + \log f(\theta)$$

we can see that prior information has les and less importance as the number of data augments.

7.2. Vague prior information

7.2.1. *A vague definition of vague prior information*

It is infrequent to find exact prior information. Usually there is prior information, but it is not clear how to formalize it in order to describe this information using a prior distribution. For example, if we are going to estimate the heritability of litter size of a rabbit breed we know that this heritability has been also estimated in other breeds and it has given often values between 0.05 and 0.11. We have a case in which the estimate was 0.30, but the standard error was high. We have also a high realized heritability in an experiment performed in Ghana, but our prejudices prevent us to take this experiment too seriously. A high heritability was also presented in a Congress, but this paper did not pass the usual peer review filter and we tend to give less credibility to this result. Moreover, some of the experiments are performed in situations that are more similar to our experiment, or with breeds that are closer to ours. It is obvious that we have prior information, but, how can we manage all of this?

One of the disappointments the student that has arrived to Bayesian inference attracted by the possibility of profiting prior information for his experiments receives is that modern Bayesians tend to avoid the use of prior information due to the difficulties of defining it properly. A solution for this problem was offered in the decade of the 30s by the British philosopher and by the Italian mathematician Bruno de Finetti, but the solution is unsatisfactory in many cases as we will see. They propose, in the words of De Finetti that "Probability does not exist". Thus what we call probability is just a state of beliefs. This definition has the advantage of including events like the probability of obtaining a 6 when throwing a dice and the probability of Scotland becoming an independent republic in this decade. Of course, in the first case we have some mathematical rules that will determine our beliefs and in the second we do not have these rules, but in both cases we can express sensible beliefs about the events. Transforming probability, which looks as a concept external to us, into beliefs, that looks like an arbitrary product of our daily mood, is a step that some scientists refuse to walk. Nevertheless, there are three aspects to consider:

1. It should be clear that although beliefs are subjective, this does not mean that they are arbitrary. Ideally, the previous beliefs should be expressed by experts and there should be a good agreement among experts on how prior information is evaluated.

2. Prior beliefs should be vague and contain little information; otherwise there is no reason to perform the experiment, as we have seen in 7.1.3. In some cases an experiment may be performed in order to add more accuracy to a previous estimation, but this is not normally the case.

3. Having data enough, prior information loses importance, and different prior beliefs can give the same result, as we have seen in 7.1.3 ([2]).

There is another problem of a different nature. In the case of multivariate analyses, it is almost impossible to determine a rational state of beliefs. How can we determine our beliefs

[2] Bayesian statisticians often stress that having data enough the problem of the prior is irrelevant. However, having data enough, Statistics is irrelevant. The science of Statistics is useful when we want to determine which part of our observation is due to random sampling and what is due to a natural law.

about the heritability of first trait, when the second trait has a heritability of 0.2, the correlation between both traits is -0.7, the heritability of the third trait is 0.1, the correlation between the first and the third traits is 0.3 and the correlation between the second and the third trait is 0.4; then our beliefs about the heritability of the first trait when the heritability of the second trait is 0.1, …etc.? Here we are unable of represent any state of beliefs, even a vague one.

7.3. No prior information

7.3.1. *Flat priors*

Since the origins of Bayesian inference (Laplace, 1774) and during its development in the XIX century, Bayesian inference was always performed under the supposition of prior ignorance represented by flat priors. Laplace himself, Gauss, Pearson and others suspected that flat priors did not represent prior ignorance, and moved to examine the properties of the sampling distribution. Integrating prior information was not proposed until the work of de Finetti quoted before.

It is quite easy to see why flat priors cannot represent ignorance: Suppose we think that we do not have any prior information about the heritability of a trait. If we represent this using a flat prior (figure 4.3), the event A "the heritability is lower than 0.5" has a probability of 50% (blue area). Take now the event B "the square of the heritability is lower than 0.25". This is exactly the same event as event A, thus its probability should be the same, also a 50%.

We are as ignorant about h^2 as about h^4, thus we should represent the ignorance about h^4 also with flat priors if they represent ignorance. However, if we do this and we also maintain that $P(h^4<0.25) = 50\%$ we arrive to an absurd conclusion: we do not know nothing about h^2 but we know that h^4 is closer to zero than h^2 (figure 4.3).

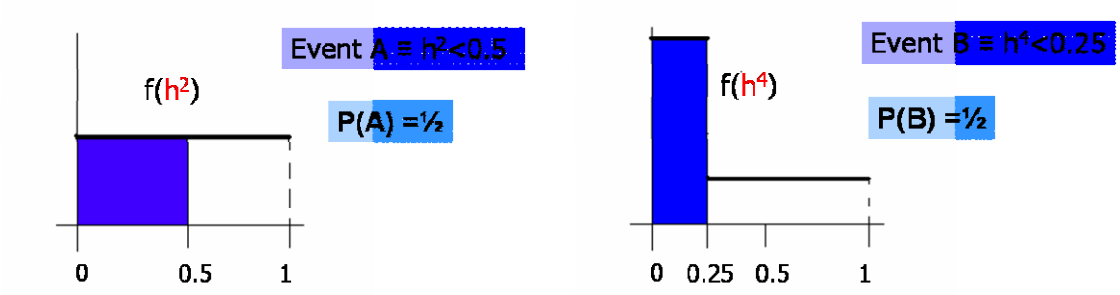

Figure 4.3. Flat priors are informative

To avoid this absurd conclusion we have to admit that flat priors do not represent ignorance, but they are informative. The problem is that we do not know what this prior information really means. However, this information is very vague and should not cause problems; in most cases that data will dominate and the results will not be practically affected by the prior.

7.4. *Improper priors*

Some priors are not densities, for example: $f(\theta) = k$, where k is an arbitrary constant, is not a density because $\int f(\theta)\, d\theta = \infty$. However, improper priors lead to proper posterior densities when

$$f(y) = \int f(y|\theta)\, f(\theta)\, d\theta < \infty$$

Sometimes they are innocuous and they are not used in the inference, for example,

y~N(μ,1)
μ ~ k

$$f(y) = \int f(y \mid \mu) f(\mu)\, d\mu = \int f(y \mid \mu) \cdot k \cdot d\mu = k \cdot \int f(y \mid \mu) \cdot d\mu =$$

$$= k \int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2} \right] d\mu = k \int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(\mu-y)^2}{2} \right] d\mu = k$$

$$f(\mu \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \mu) f(\mu)}{f(\mathbf{y})} = \frac{f(\mathbf{y} \mid \mu) \cdot k}{k} = f(\mathbf{y} \mid \mu)$$

thus in this case the posterior density of μ does not take into account the prior.

In general, it is recommended to use always proper priors, to be sure that we always obtain proper posterior densities. When using Gibbs sampling, some densities look as proper ones and they may be improper. Although when using MCMC all densities are in practice proper ones (we never sample in the infinite), samples can have very long burning periods and can lead to chains that only apparently have converged. The recommendation is always to used proper priors (bounded priors with reasonable limits, for example), unless it has been proved that they are innocuous (Hobert and Casella, 1992).

7.5. *The Achilles heel of Bayesian inference*

Bayesian inference, or Inverse probability, as it was always called before and should still be called, is extremely attractive because of the use of probabilities and the possibility of integrating prior information. However, integrating prior information is much more difficult that the optimistic Bayesians of the fifties thought. This lead to use several artefacts in order to make possible the use of probability. Some statisticians think that an artefact multiplied by a probability will give an artefact and not a probability, and consequently they are reluctant to use Bayesian inference. There is not a definitive answer to this problem, and it is a matter of opinion to use Bayesian or frequentist statistics, both are now widely used and no paper will be refused by a publisher because it uses a type or the other type of statistics.

Many users of statistics, like the author of this book, are not “Bayesians” or “frequentists”, but just people with problems. Statistics is a tool to help in solving these problems, and users depend more on the existence of easy solutions and friendly software than in the background philosophy. I use Bayesian statistics because I understand probability better than significance levels and because it permits to me to express my results in a more clear way for later discussion. Some other users prefer Bayesian statistics because there is a route for solving their problems: to make a joint posterior distribution, to find the conditionals and to use MCMC to find the marginal distributions. We *behave as* if we were working with real probabilities, (this should not be objected by frequentists). To know the true probabilities drives us to the PROBLEM OF INDUCTION, a very difficult problem that we cannot expose in this lecture notes.