5. The prior distribution

Objective

Firstly, we study the different methods for elicitation, calibration and combination of *subjective* (expert) prior distributions and secondly, we analyze the different *objective* Bayesian approaches.

Recommended reading

- Berger, J. (2006). The case for objective Bayesian analysis (with discussion). *Bayesian Analysis*, **1**, 385–482.

- Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice (with discussion). *Bayesian Analysis*, **1**, 403–420.

Both articles available from http://ba.stat.cmu.edu/vol01is03.php

Subjective probability distributions and their elicitation

In many real problems, substantive experts with important information are available. In such cases, it is important to be able to convert their information and ideas into probabilities. The techniques for doing this are called *probability elicitation* methods.

Garthwaite et al (2005) observe that there should be four stages in any probability elicitation exercise:

- Setting up: selection and training of experts, investigation of parameters to make judgements about.
- Elicitation of expert opinions: summaries of the experts probability distributions for some parameter.
- Distribution fitting: combining and modeling the expert opinions.
- Feedback: presentation of results to experts to see if they agree or not.

Expert training: problems with the use of expert judgements

Experts often use heuristic methods to make their probability judgements which can induce biases and incoherence.

- motivational biases.
- cognitive biases:
  - ◇ availability
  - ◇ anchoring
  - ◇ representativeness
  - ◇ control

An important part of the training is to try to get the experts to recognize such biases so that they may be eliminated.

Example 22

For each of the following pairs, which causes more deaths per year.

- Stomach cancer or car accidents?
- Tuberculosis or fires?

Cause of death	Choice	Total yearly death rate (USA /1000)	# newspaper articles
cancer	14%	95	46
car accidents	86%	1	137
tuberculosis	23%	4	0
fires	77%	5	0

People have much more information about accidents than about cancer and therefore, this option is more available. See Russo and Shoemaker (1989)

See Tversky and Kahneman (1973) for more examples of the availability bias.

Example 23
(Tversky y Kahneman 1974).

The researcher wished to elicit an estimation of the percentage of African countries in the UN. One group of experts were asked the question

Do you think that the percentage is higher or lower than 10%?

and the second group were asked

Do you think that the percentage is higher or lower than 65%?

The members of both groups were then asked to give a point estimate of the percentage.

The mean estimate in group 1 was 25% and the mean in group 2 was 45%. A random and irrelevant anchoring point has influenced the estimates of both groups.

Example 24

Federico is 35 years old, intelligent but not very imaginative and a bit boring. In college, he showed a lot of talent in maths but he wasn't very good at art.

Order the following statements about Federico in terms of their probability (1 = most probable, 8 = least probable).

1. Federico is a doctor and likes to play cards as a hobby.
2. He is an architect.
3. He is an accountant.
4. He plays a jazz instrument.
5. He reads *Marca*.
6. He likes mountaineering.
7. He is an accountant and plays a jazz instrument.
8. He is a journalist.

1. Federico is a doctor and likes to play cards as a hobby.
2. He is an architect.
3. He is an accountant.
4. He plays a jazz instrument.
5. He reads $Marca$.
6. He likes mountaineering.
7. He is an accountant and plays a jazz instrument.
8. He is a journalist.

Most people say that option 3 is the most probable. Moreover, many people say that option 7 is more probable than option 4. This is impossible as for any two events A and B,

$$P(A \cap B) \leq \min\{P(A), P(B)\}.$$

This problem illustrates the representativeness heuristic and also the base rate fallacy. See Kahneman et al (1982) for more examples.

The base rate fallacy

Example 25
(Tversky and Kahneman 1980).

A taxi knocks over a pedestrian in Darlington. In Darlington, only two companies operate taxi services. The first company has green taxis and the second operates blue taxis. Around 85% of the taxis in Darlington are green.

There is a witness to the accident who says that the taxi was blue. When tested under the same climatological conditions as the night of the accident, the witness identifies the two colours correctly in around 80% of the test cases.

What is your estimated probability that the taxi is blue?

The typical response is around 80%. However, if A represents the event that the taxi is blue and a is the event that the witness says it is blue, then from Bayes Theorem:

$$\begin{aligned} P(A|a) &= \frac{P(a|A)P(A)}{P(a|A)P(A) + P(a|\bar{A})P(\bar{A})} \\ &= \frac{0.8 \times 0.15}{0.8 \times 0.15 + 0.2 \times 0.85} = 0.41 \end{aligned}$$

People often ignore the base rate when making their predictions.

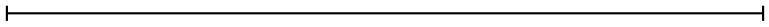
Elicitation

Initially we shall consider the simplest problem of eliciting an expert, E's probability, $P_E(A)$, that an event occurs. The simplest approach might appear to be to ask the expert directly for her probability. However, this method does not allow her to think about her probabilities and if she is not statistically trained, will be very hard to implement.

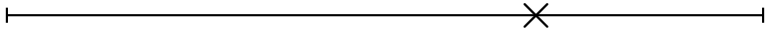
There are two basic alternative approaches based on the use of probability scales or on gambling schemes.

Probability scales

The simplest form of probability scale is a just a straight, unmarked line where the right hand end indicates probability one and the left, probability zero.



The expert is asked to mark a point which represents her probability of a given event.

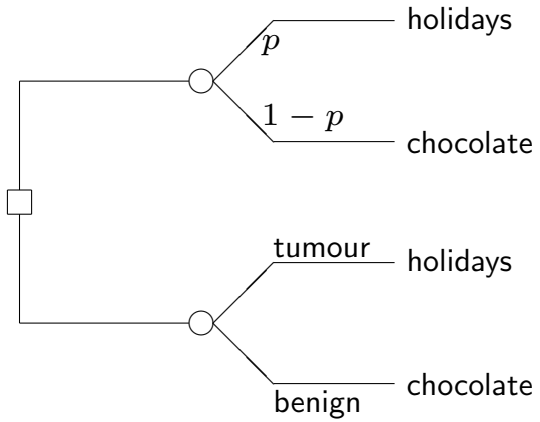


The simple probability scale will not allow the expert to estimate small or large probabilities well. When we are estimating such probabilities, it is better to use an odds scale or a logarithmic scale. Also, it is often useful to include certain guide points on the scale although these might induce *anchoring biases*.

One typical approach is to use a lottery. Recall, from page 30, that De Finetti (1937) defined subjective probabilities in terms of (*certainty equivalent*) lotteries. Another approach is to consider gambles with one big prize and one negligible prize.

Example 26

Suppose that we wish to elicit a doctor's probability, p_E, that a given patient has a tumour. Then we might ask the doctor to choose to play one of two lotteries as in the following diagram.



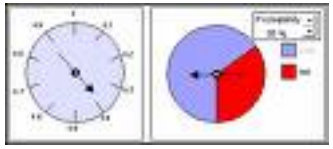
For a given value of p we can check which lottery is preferred and then we can vary p until the doctor is indifferent at some point $p = p_E$. Then $P(\text{tumour}) = p_E$ is the doctor's elicited probability that the patient has the tumour.

There are certain ethical problems with the use of lotteries like this. An alternative is to use *probability wheels*.

Example 27

The illustration shows the *Spinner* probability wheel from *Insight*©.

http://www.stanford.edu/~savage/faculty/savage/InsightInfo.xls



We now ask the doctor to say which event is more likely: that the pointer lands in red or that the patient has a tumour. By varying the size of the coloured sectors we can arrive at a point of indifference when $p = p_E$ is the proportion of the disc coloured red.

Both probability wheel and lottery approaches are basically restricted to assessment of binary probabilities. Also, neither method will work very well if we wish to estimate very small or very large probabilities.

Alternative approaches have been developed based on frequencies (Price 1998) or on attempting to translate verbal expressions such as likely, improbable, etc., into numerical probabilities (Witteman and Renooij 2003). See e.g. Wallsten et al (1993) for a fuller review.

These methods can be extended to elicitation of expert distributions. In these cases, it is most common to elicit $quantiles$ from the expert rather than the full distribution function. These quantiles can then be fitted to a given distributional model as required.

There are many possibilities. The worst would be to ask the expert directly about the parameters of the prior distribution.

Example 28

Suppose we are interested in estimating a prior distribution for the probability, p, of heads for a biased coin. In this case, we know that the conjugate prior is beta, $p \sim \mathcal{B}(\alpha, \beta)$, and we wish to derive the values of α and β.

One possibility (Fox 1966) is to ask the expert for a direct estimate of the most likely value of p, i.e. the expert's mode, say p_E, and to state the probability, r_E, that the true value of p lies in an interval $(p_E - Kp_E, p_E + Kp_E)$ where the value of K is fixed by the analyst eliciting the information. Then, assuming that p has a beta prior, $p \sim \mathcal{B}(\alpha, \beta)$, then we can find the values of α and β best representing the expert's judgements by solving

$$
\begin{aligned}
p_E &= \frac{\alpha - 1}{\alpha + \beta - 2} \\
r_E &= \int_{p_E - Kp_E}^{p_E + Kp_E} \frac{1}{B(\alpha, \beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} \, dx
\end{aligned}
$$

In general, it is preferable to ask the expert about $observable$ quantities.

Chaloner and Duncan (1983) propose asking the expert to state her mode, x_E, for the number of successes, X, that would occur in a given number, n, of Bernoulli trials and how much less likely it would be that that the number of successes is one less or one more than the mode, say

$$c_E = \frac{P_E(X = x_E - 1)}{P_E(X = x_E)} \quad d_E = \frac{P_E(X = x_E - 1)}{P_E(X = x_E)}.$$

Then, recalling that the marginal distribution of X supposing a beta prior is beta-binomial, the parameters may be estimated by solving the following system of equations for α and β.

$$c_E = \frac{(n - x_E)(x_E + \alpha)}{(x_E + 1)(n - x_E + \beta - 1)} \quad \text{and} \quad d_E = \frac{x_E(n - x_E + \beta)}{(n - x_E + 1)(n - \alpha + 1)}.$$

Many other approaches have been considered. See e.g. Hughes and Madden (2002).

Evaluating the quality of expert forecasts

Typically used criteria (Lichtenstein et al 1982) are the following:

- *Honesty*: We want the expert to tell the truth.
- *Coherence*: Her forecasts should satisfy the laws of probability.
- *Consistency*: If she doesn't receive new information, then her predictions shouldn't change.
- *Calibration*: It should rain on around 50% of the days when the expert says $P(\text{rain}) = 0.5$.
- *Informativeness:* If, in Madrid, it rains on around 50 days per year, an expert who says

  $$P(\text{rain tomorrow}) = 50/364$$

  every day isn't very informative.

Honesty and strictly proper scoring rules

Suppose that we wish to elicit the expert's true probability, p_E, that some event A occurs. One method of encouraging the expert to be honest is to pay her a quantity $R(A, p)$ which depends upon the occurrence or not of A and the expert's stated probability, p.

How should we define $R(A, p)$?.

We suppose that the expert wishes to maximize her expected income. If p_E is her true probability, her expected income if she states a probability p is

$$p_E R(1, p) + (1 - p_E) R(0, p).$$

Definition 9

A *(strictly) proper scoring rule* (Savage 1971) is a scoring rule $R(A, p)$ whereby the expert maximizes his expected income if (and only if) $p = p_E$.

Example 29

Suppose that $R(A,p) = 1 - |A - p|$. Then, the expert's expected earnings if she states a probability p are

$$\begin{aligned} E[R] &= p_E\left(1 - |1-p|\right) + (1-p_E)\left(1 - |0-p|\right) \\ &= p_E p + (1-p_E)(1-p) \\ &= 1 - p_E + (2p_E - 1)p \end{aligned}$$

Therefore $R(A,p)$ is not a proper scoring rule as the expert maximizes her expected earnings by stating $p = 1$ (0) if $p_E > (<) 0.5$.

Example 30

$R(A, p) = 1 - (A - p)^2$ is the Brier (1950) score.

$$\begin{aligned} E[R] &= p_E \left(1 - (1-p)^2\right) + (1 - p_E)\left(1 - p^2\right) \\ &= 1 - p_E + 2pp_E - p^2 \\ &= 1 - p_E + p_E^2 - (p - p_E)^2 \end{aligned}$$

which is maximized by setting $p = p_E$. Therefore, R is a strictly proper scoring rule.

There are many other proper scoring rules, see e.g. Winkler (1986) and proper scoring rules have also been developed for continuous variables and quantiles. See e.g. Buehler (1971) and Matheson and Winkler (1976).

Example 31

Suppose that E is asked to state a point estimator, say e, for a variable X. Then consider the scoring rule

$$R(X,e) = \begin{cases} a(e-x) & \text{if } e < x \\ b(x-e) & \text{if } e > x \end{cases}$$

Let $p_E(x)$ be the expert's true distribution for X:

$$\begin{aligned} E[R(X,e)] &= \int R(x,e)p_E(x)\,dx \\ &= a\int_e^\infty (e-x)p_E(x)\,dx + b\int_{-\infty}^e (x-e)p_E(x)\,dx \\ &= ae(1-F_E(e)) - a\int_e^\infty xp_E(x)\,dx + \\ &\quad b\int_{-\infty}^e xp_E(x)\,dx - beF_E(e) \end{aligned}$$

$$
\begin{aligned}
\frac{dE[R(X,e)]}{de} &= a(1-F_E(e)) - aep_E(e) + aep_E(e) - \\
&\quad bep_E(e) - bF_E(e) + bep_E(e) \\
0 &= a(1-F_E(\hat{e})) - bF_E(\hat{e}) \\
F_E(\hat{e}) &= \frac{a}{a+b}
\end{aligned}
$$

The expert maximizes her experted gains if she states her $b/(a+b) \times 100\%$ quantile. See Raiffa and Schlaifer (1961).

The use of proper scoring rules to encourage honesty seems somewhat artificial. However, they may also be used $a\ posteriori$ as evaluation tools for expert probabilities.

Numerical measures of expert quality

Suppose that the expert supplies probabilities $\mathbf{p}_E$ for a sequence of Bernoulli events $X_1, \ldots, X_n$. Given the data $\mathbf{x}$, we wish to evaluate the quality of her predictions.

Consider the Brier score

$$R(X, p_E) = 1 - (X - p_E)^2.$$

Then, given the data, we can calculate the statistic

$$R(\mathbf{x}, \mathbf{p}_E) = \frac{1}{n} \sum_{i=1}^{n} R(x_i, p_{Ei}) = 1 - \sum_{i=1}^{n} (x_i - p_{Ei})^2$$

which is a measure of the average quality of her predictions.

This measure can be divided into a measure of calibration and a measure of information.

Following Murphy (1973), assume that the expert uses the probability p_j a total of n_j times, with a frequency of f_j successes and a relative frequency of $r_j = f_j/n_j$ successes for $j = 1, \ldots k$. Then:

$$\begin{aligned} R(\mathbf{x}, \mathbf{p}) &= 1 - \frac{1}{n} \sum_{j=1}^{k} \left(f_j (1 - p_j)^2 + (n_j - f_j)(0 - p_j)^2 \right) \\ &= 1 - \frac{1}{n} \sum_{j=1}^{k} n_j \left(r_j (1 - p_j)^2 + (1 - r_j)(0 - p_j)^2 \right) \end{aligned}$$

Now we can prove the following theorem.

Theorem 25

$R(\mathbf{x}, \mathbf{p}) = 1 - C(\mathbf{x}, \mathbf{p}) - I(\mathbf{x}, \mathbf{p})$ where

$$C(\mathbf{x}, \mathbf{p}) = \frac{1}{n} \sum_{j=1}^{k} n_j (r_j - p_j)^2 \quad \text{is a measure of calibration}$$

$$I(\mathbf{x}, \mathbf{p}) = \frac{1}{n} \sum_{j=1}^{k} n_j r_j (1 - r_j) \quad \text{is a measure of information}$$

Proof

$$
\begin{aligned}
R(\mathbf{x},\mathbf{p}) &= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j(1-p_j)^2+(1-r_j)(0-p_j)^2\right)\\
&= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j-2r_jp_j+p_j^2\right)\\
&= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j-r_j^2+r_j^2-2r_jp_j+p_j^2\right)\\
&= 1-\frac{1}{n}\sum_{j=1}^{k} n_j\left(r_j(1-r_j)+(r_j-p_j)^2\right)\\
&= 1-I(\mathbf{x},\mathbf{p})-C(\mathbf{x},\mathbf{p})
\end{aligned}
$$

C has the following properties:

- $0 \leq C \leq 1$
- $C = 0$ if and only if $r_j = p_j$ for $j = 1, \ldots, k$.
- For a well calibrated expert, when $n \to \infty$, $C \to 0$.
- C is large if the observed relative frequencies r_i are very different from the expert's stated probabilities p_i.

I has the following properties:

- $0 \leq I \leq 0.25$.
- $I = 0$ if, for all p_j, the relative frequency $r_j = 0$ or 1.
- $I = 0.25$ if, for all p_j, then $r_j = 0.5$.

Any strictly proper scoring rule can be divided up into calibration and information measures in a similar way. See e.g. De Groot and Fienberg (1982).

Example 32

Wiper (1987,1990) gave 12 experts a set of 50 statements to study. The experts were asked to state whether each statement was true or false and to give their probabilities that they were correct as a percentage between 50% (= no idea) and 99%.

We shall assume here that the experts' stated probabilities for all events belong to the class $\mathbf{p} = \{0.53, 0.64, 0.75, 0.86, 0.97\}$. Then the following table gives the related absolute and relative frequencies.

E	p_i	0.53	0.64	0.75	0.86	0.97
2	n_i	25	6	6	5	8
	f_i	15	4	4	3	8
	r_i	0.6	0.67	0.67	0.6	1.0
3	n_i	25	5	10	5	5
	f_i	16	1	3	2	4
	r_i	0.64	0.2	0.3	0.4	0.8
10	n_i	10	5	15	1	19
	f_i	6	2	5	0	15
	r_i	0.6	0.4	0.3	1.0	0.79

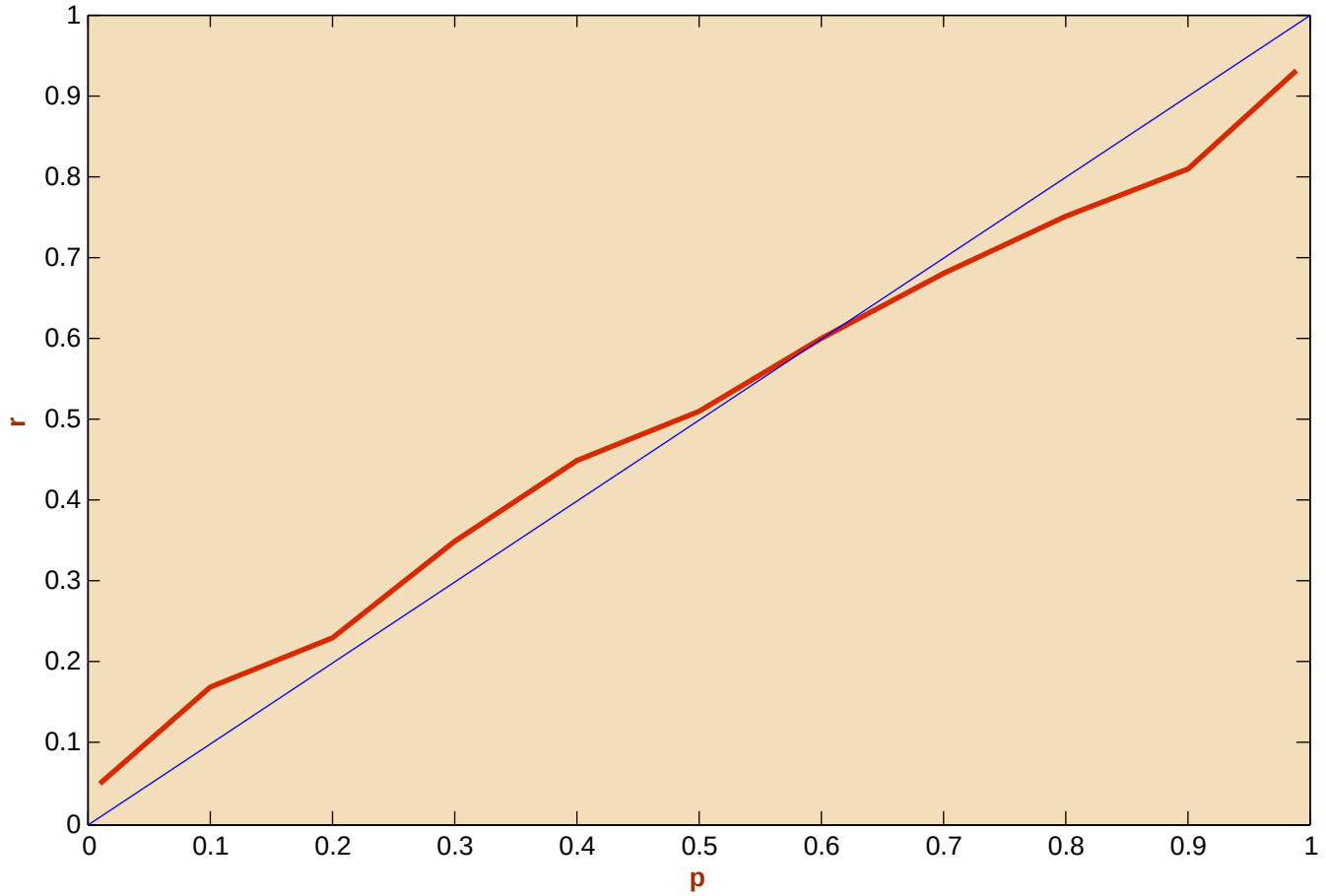
The following table shows the calibration, information and Brier scores for each expert.

E	C	I	Brier
2	.0093	.1973	.7934
3	.0900	.2132	.6968
10	.1059	.2018	.6923

We can see that expert 2 is better calibrated but less informative than the other experts.

A visual manner of illustrating expert calibration is provided by the calibration curve. This is simply a graph of observed frequencies, r_j, against the different probabilities used, p_j.

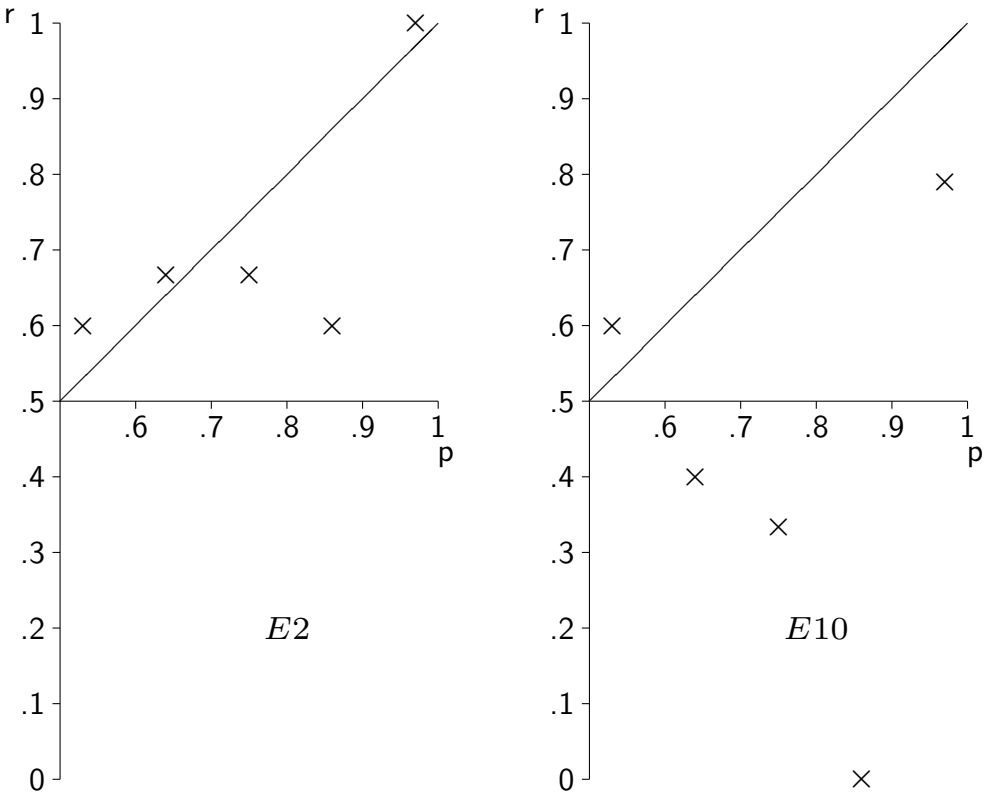
The calibration curve



For a well calibrated expert, the curve approximates the 45 degree line.

Example 33

Returning to Example 32, the calibration curves of experts 2 and 10 are given in the following figure.



Cooke's approach

Cooke et al (1988) and Cooke (1991) have developed alternative measures of calibration and information based on classical p-values which can be applied to predictions for both discrete and continuous variables.

Suppose that an expert uses the probability p_j a total of n_j times for $j = 1, 2, \ldots, k$. Theoretically, if the expert is well calibrated, about the total frequency of events $\{X = 1\}$ that occur should be around $n_j \times p_j$.

A χ^2 test can be set up to test whether the observed relative frequencies $r_1, \ldots, r_k$ could be generated from the theoretical distribution $p_1, \ldots, p_k$.

Thus, to test $H_0 : r \sim p$ against the alternative $H_1 : r \nsim p$ we calculate the chi-squared statistic

$$S = \sum_{j=1}^{k} n_j \frac{(r_j - p_j)^2}{r_j}.$$

The hypothesis that the expert is well calibrated can then be accepted or rejected by comparing S with tables of the χ^2_k distribution.

Example 34

The following table shows the p-values for each expert in Example 32.

E	2	3	10
p	.7	.00	.00

It seems that only expert 2 is reasonably well calibrated.

Cooke (1991) derives a theory of scoring rules based on combining the p-value with a measure of information. When the expert makes forecasts for continuous variables, then an alternative is a Kolmogorov Smirnov test. See Wiper et al (1994).

Objective Bayesian methods

Sometimes we wish to use a prior distribution which does not include (much) subjective information, because

- we don't know anything about the problem at hand,
- we would like to be *objective*.

In such situations, we should choose a *non-informative* prior distribution.

However, there are many possible elections. Which is the most useful?

Uniform priors

Bayes (1763) and Laplace (1812) generally employed uniform prior distributions, as justified by the principle of insufficient reason, see page 23. This is fine in finite dimensional problems, but if the parameter space Θ is continuous or uncountable, then such prior distributions are $improper$. In practice, this is not a serious problem, as long as the posterior distribution as calculated via Bayes theorem can be shown to exist.

An important problem with the general use of uniform priors however is lack of invariance to transformation.

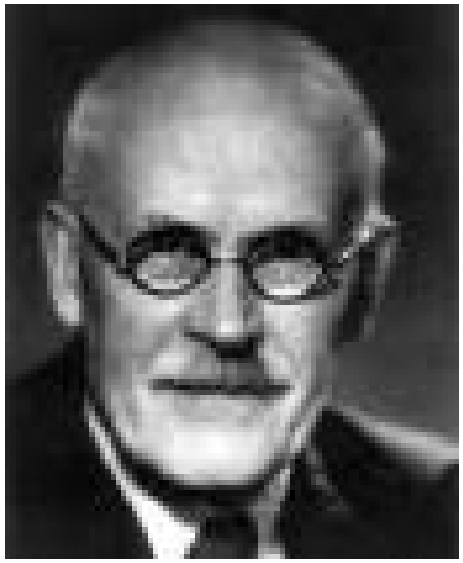
Example 35

Suppose that we set $p(\theta) \propto 1$ and define the transformed variable $\phi = \frac{1}{\theta}$. Then the implied prior distribution for ϕ is

$$p(\phi) \propto \frac{1}{\phi^2}$$

which is not uniform. Thus, the use of the uniform distribution to represent lack of knowledge is inconsistent. If we know nothing about θ, then we should know nothing about ϕ.

Jeffreys priors



Jeffreys

Jeffreys (1946) introduced a prior distribution which possesses an invariance property.

Definition 10

Let $X|\theta \sim f(\cdot|\theta)$ where θ is one dimensional. The Jeffreys prior for θ is

$$p(\theta) \propto \sqrt{I(\theta)}$$

where $I(\theta) = -E_X\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right]$ is the expected Fisher information.

The following theorem shows that if the Jeffreys prior for θ is used, then the implied prior for the transformed parameter $\phi = \phi(\theta)$ is the Jeffreys prior for ϕ.

Theorem 26

If $\phi = \phi(\theta)$, and $p(\theta)$ is the Jeffreys prior for θ, then the implied prior for ϕ is the Jeffreys prior

$$p(\phi) \propto \sqrt{I(\phi)}$$

where $I(\phi) = -E_X\left[\frac{d^2}{d\phi^2}\log f(X|\phi)\right]$ is the expected Fisher information when the distribution of X is reparameterized in terms of ϕ.

Proof Firstly, we shall prove that $E\left[\frac{d}{d\theta}\log f(X|\theta)\right] = 0$.

$$\begin{aligned} E_X\left[\frac{d}{d\theta}\log f(X|\theta)\right] &= \int\left[\frac{d}{d\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\ &= \int \frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\,dx \\ &= \int f'(x|\theta)\,dx \\ &= \frac{d}{d\theta}\int f(x|\theta)\,dx = 0. \end{aligned}$$

Secondly we shall prove that

$$E\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] = -E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right].$$

We have

$$
\begin{aligned}
E_X\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] &= \int \frac{d^2}{d\theta^2}\log f(x|\theta)\,dx \\
&= \int \frac{d}{d\theta}\left[\frac{d}{d\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\
&= \int \frac{d}{d\theta}\frac{f'(x|\theta)}{f(x|\theta)} f(x|\theta)\,dx \\
&= \int \left[\frac{f''(x|\theta)}{f(x|\theta)} - \left(\frac{f'(x|\theta)}{f(x|\theta)}\right)^2\right] f(x|\theta)\,dx \\
&= \int \frac{d^2}{d\theta^2} f(x|\theta)\,dx - \int \left(\frac{d}{d\theta}\log f(x|\theta)\right)^2 f(x|\theta)\,dx \\
&= -E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]
\end{aligned}
$$

Now finally, we have that

$$
\begin{aligned}
p(\phi) &= p(\theta)\left|\frac{d\theta}{d\phi}\right| \quad \text{and squaring,} \\
p(\phi)^2 &= p(\theta)^2\left|\frac{d\theta}{d\phi}\right|^2 \\
&\propto -E\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]\left|\frac{d\theta}{d\phi}\right|^2 \\
&\propto E\left[\left(\frac{d}{d\phi}\log f(X|\phi)\right)^2\right] \\
p(\phi) &\propto \sqrt{I(\phi)}.
\end{aligned}
$$

Example 36

$X|\theta \sim \mathcal{BI}(n, \theta)$.

$$\begin{aligned}
\log f(X|\theta) &= c + X \log \theta + \\
&\quad +(n - X) \log(1 - \theta) \\
\frac{d}{d\theta} \log f(X|\theta) &= \frac{X}{\theta} - \frac{(n - X)}{(1 - \theta)} \\
\frac{d^2}{d\theta^2} \log f(X|\theta) &= -\frac{X}{\theta^2} - \frac{(n - X)}{(1 - \theta)^2} \\
E\left[\frac{d^2}{d\theta^2} \log f(X|\theta)\right] &= -n\left(\frac{1}{\theta} + \frac{1}{1 - \theta)}\right) \\
I''(\theta) &\propto \frac{1}{\theta(1 - \theta)}
\end{aligned}$$

Therefore, the Jeffreys prior is $p(\theta) \propto \sqrt{\frac{1}{\theta(1-\theta)}}$, that is $\theta \sim \mathcal{B}(1/2, 1/2)$. This is a *proper* prior, unlike Haldane's prior which we saw earlier.

Example 37

$X|\mu \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$, with ϕ known.

$$\begin{aligned} \log f(X|\mu) &= c - \frac{\phi}{2}(X-\mu)^2 \\ \frac{d}{d\mu}\log f(X|\mu) &= \phi X - \phi\mu \\ \frac{d^2}{d\mu^2}\log f(X|\mu) &= -\phi \end{aligned}$$

We have $p(\mu) \propto 1$, a uniform distribution.

Example 38

Suppose now that $X|\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$ where μ is known.

$$\begin{aligned}
\log f(X|\phi) &\propto \frac{1}{2}\log\phi - \phi\frac{(X-\mu)^2}{2} \\
\frac{d}{d\phi}\log f(X|\phi) &= \frac{1}{2\phi} + \frac{(X-\mu)^2}{2} \\
\frac{d^2}{d\phi^2}\log f(X|\phi) &= -\frac{1}{2\phi^2}
\end{aligned}$$

and the Jeffreys prior for ϕ is

$$p(\phi) \propto \frac{1}{\phi}.$$

If σ is the standard deviation, then $\phi = \frac{1}{\sigma^2}$ and $\frac{d\phi}{d\sigma} = -\frac{2}{\sigma^3}$ so the Jeffreys prior for σ is

$$p(\sigma) \propto \sigma^2\frac{2}{\sigma^3} \propto \frac{1}{\sigma}.$$

Jeffreys priors in multivariate problems

It is possible to extend the definition of a Jeffreys prior to the case when $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is multivariate, by defining $p(\boldsymbol{\theta}) \propto \sqrt{I(\boldsymbol{\theta})}$ as earlier, where the expected Fisher information is now given by

$$I(\boldsymbol{\theta}) = |E_X [\mathbf{J}(\boldsymbol{\theta})]| \quad \text{where}$$

$$\mathbf{J}_{ij} = \frac{d^2}{d\theta_i d\theta_j} \log f(X|\boldsymbol{\theta}).$$

In some cases, the multivariate Jeffreys prior seems reasonable.

Example 39

Let $X|\boldsymbol{\theta} \sim \mathcal{MN}(m, \boldsymbol{\theta})$ have a (k dimensional) multinomial distribution. Then

$$\begin{aligned}
\log f(X|\boldsymbol{\theta}) &= c + \sum_{i=1}^{k} x_i \log \theta_i \\
\frac{d}{d\theta_i} \log f(X|\boldsymbol{\theta}) &= \frac{X_i}{\theta_i} \\
\frac{d^2}{d\theta_i^2} \log f(X|\boldsymbol{\theta}) &= -\frac{X_i}{\theta_i^2} \\
\frac{d^2}{d\theta_i d\theta_j} \log f(X|\boldsymbol{\theta}) &= 0 \quad \text{for } i \neq j. \\
E\left[\frac{d^2}{d\theta_i^2} \log f(X|\boldsymbol{\theta})\right] &= -\frac{m}{\theta_i} \\
I(\boldsymbol{\theta}) &= \frac{m^k}{\prod_{i=1}^{k} \theta_i}
\end{aligned}$$

and the Jeffreys prior is $p(\boldsymbol{\theta}) \propto \prod_{i=1}^{k} \frac{1}{\sqrt{\theta_i}}$ which is Dirichlet, $\boldsymbol{\theta} \sim \mathcal{D}\left(\frac{1}{2}, \ldots, \frac{1}{2}\right)$.

In many more cases, the multivariate Jeffreys prior is less natural.

Example 40

Let $X|\mu,\phi \sim \mathcal{N}\left(\mu, \frac{1}{\phi}\right)$. Then, from Examples 37 and 38,

$$\begin{aligned}
\frac{d^2}{d\mu^2}\log f(X|\mu,\phi) &= -\phi \\
\frac{d^2}{d\phi^2}\log f(X|\mu,\phi) &= -\frac{1}{2\phi^2} \quad \text{and also} \\
\frac{d^2}{d\mu d\phi}\log f(X|\mu,\phi) &= -(X-\mu) \\
E[\mathbf{J}] &= -\begin{pmatrix} \phi & 0 \\ 0 & \frac{1}{2\phi^2} \end{pmatrix}
\end{aligned}$$

and therefore, the Jeffreys prior is $p(\mu,\phi) \propto \frac{1}{\sqrt{\phi}}$ which is not the prior we have used earlier.

Maximum entropy priors

Jaynes

The idea (Jaynes 1968,1983) is to find the least informative prior distribution in the presence of partial information.

Entropy

Assume that θ is univariate and discrete. If $p(\theta)$ is any distribution for θ, then we can define

$$e(p) = -\sum_{i \in \Theta} p(\theta_i) \log p(\theta_i)$$

to be the $entropy$ of the distribution.

If $p(\theta = \theta_i) = 1$ for some value $\theta_i \in \Theta$, then, $e(p) = 0$ and there is zero uncertainty or minimum entropy. On the contrary, if $p(\theta_i) = 1/|\Theta|$, i.e. a uniform distribution, then

$$e(p) = -\sum_{i \in \Theta} \frac{1}{|\Theta|} \log \frac{1}{|\Theta|} = \log |\Theta|,$$

the maximum entropy.

Maximum entropy (maxent) distributions are minimum information distributions.

In many practical situations, we may only wish to fix certain characteristics of the prior distribution, e.g. quantiles or moments and apart from this, let the prior be as uninformative as possible.

Suppose that we have partial information about θ in the form

$$E[g_k(\theta)] = \sum_{i \in \Theta} p(\theta_i) g_k(\theta_i) = \mu_k$$

for $k = 1, \ldots, m$. This includes fixed moments, e.g. $g_1(\theta) = \theta$ and quantiles

$$g_k(\theta) = I_{(-\infty, z_k]} \Rightarrow E[g_k(\theta)] = p(\theta \leq z_k).$$

Given these restrictions, the following theorem provides the form of the maximum entropy prior.

Theorem 27
Given the partial information

$$E[g_k(\theta)] = \sum_{i \in \Theta} p(\theta_i) g_k(\theta_i) = \mu_k$$

for $k = 1, \ldots, m$, then the maxent prior is

$$p(\theta_i) = \frac{\exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_i)\right)}{\sum_{j \in \Theta} \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta_j)\right)}$$

where the constants λ_k can be determined from the information.

Proof See Jaynes (1968). ∎

Example 41

Let $X|N \sim \mathcal{BI}(N, 1/2)$. Suppose that we know that $N \geq 1$ and that we fix the mean to be $E[N] = 10$.

We shall try to calculate the maxent distribution for N.

$$\begin{aligned} P(N = n) &= \frac{\exp(\lambda_1 n)}{\sum_{j=1}^{\infty} \exp(\lambda_1 j)} \\ &= \exp(\lambda_1 n) \frac{1 - \exp(\lambda_1)}{\exp(\lambda_1)} \\ &= (1 - e^{\lambda_1}) \exp(\lambda_1 (n-1)) \end{aligned}$$

Thus, $N - 1$ has a geometric density with parameter $1 - e^{\lambda_1}$ and therefore

$$E[N] = 1 + \frac{e^{\lambda_1}}{1 - e^{\lambda_1}}$$

and fixing $E[N] = 10$, we have $e^{\lambda_1} = \frac{9}{10}$, i.e. the maxent prior for N is $N - 1 \sim \mathcal{GE}(9/10)$.

Maxent priors for continuous variables

The extension to continuous variables is more complicated because the definition of entropy

$$e(p) = -\int p \log p \, d\mu$$

depends on the base measure μ.

One possibility (Jaynes 1968) is to define

$$e(p) = -\int p(\theta) \log \frac{p(\theta)}{p_0(\theta)} \, d\theta$$

where $p_0(\theta)$ is the Jeffreys prior for θ. Then, given the restrictions $E[g_k(\theta)] = \lambda_k$, the maxent prior is

$$p(\theta) = \frac{p_0(\theta) \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta)\right)}{\int p_0(\theta) \exp\left(\sum_{k=1}^{m} \lambda_k g_k(\theta)\right) \, d\theta}$$

analogous to the discrete case.

Unfortunately, in some cases, it is possible that no maxent prior distribution exists.

Example 42

Let $X|\mu \sim \mathcal{N}(\mu, 1)$ and suppose that we fix the prior mean to be $E[\mu] = m$. We have seen in Chapter 4 that the Jeffreys prior is $p(\mu) \propto 1$. Therefore, the maxent distribution is

$$p(\mu) = \frac{\exp(\lambda_1 \mu)}{\int_{-\infty}^{\infty} \exp(\lambda_1 \mu) d\mu}$$

and there is no solution to this integral.

Reference priors

Bernardo

This, the most general approach, was developed by Bernardo (1979). It is based on maximizing the expected information about $\boldsymbol{\theta}$ to be provided by a given experiment.

Consider first the case when θ is one dimensional.

Then, given the prior distribution $p(\theta)$, the expected information about θ to be gained by observing a sample $\mathbf{X}$ of size n, where $X|\theta \sim f(\cdot|\theta)$ is defined by

$$I(p(\theta)) = \int f(\mathbf{x}) \int p(\theta|\mathbf{x}) \log \frac{p(\theta|\mathbf{x})}{p(\theta)} \, d\theta \, d\mathbf{x}$$

where $f(\mathbf{x}) = \int f(\mathbf{x}|\theta) p(\theta) \, d\theta$ and $p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)p(\theta)}{f(\mathbf{x})}$.

Then, the *reference prior* is defined to be the prior $p(\theta)$, within the class of admissible priors, which maximizes the asymptotic limit of the expected information $I(p(\theta))$ as the sample size n goes to infinity.

It can be shown that the maximum entropy and Jeffreys priors correspond to particular cases of reference priors.

Reference priors in multivariate problems

Suppose that $\boldsymbol{\theta} = (\theta_1, \theta_2)$, where θ_1 is the parameter of interest and θ_2 is a nuisance parameter. Then the reference prior approach has two steps:

1. Calculate the reference prior $p(\theta_2|\theta_1)$ as above.

2. If this is proper, then θ_2 can be integrated out of the density of X and the reference prior of θ_1 can be found as above. If not, then the procedure can be performed in a limiting way. See Bernardo (1979).

Other non-informative prior distributions

There are a number of other approaches to defining non-informative priors:

- Limiting forms of conjugate priors.

  This is the method we used in chapters 2 and 3.

- Priors based on the data translated likelihood. Box and Tiao (1973).

  This approach shows how to define the transformation $\phi = \phi(\theta)$ so that a uniform prior for ϕ is justified. The resultant distributions are Jeffreys priors.

- Methods based on symmetry, e.g. Haar priors.

- Others: see Yang and Berger (1997) and Kass and Wassermann (1996).

Problems with the use of non-informative prior distributions

There are various theoretical and practical difficulties with the use of non-informative priors. Firstly, when improper prior distributions are used, it is important to show that the posterior distribution is proper.

Example 43

Let $X|\theta \sim \mathcal{BI}(n, \theta)$ and suppose we use Haldane's prior $p(\theta) \propto \frac{1}{\theta(1-\theta)}$. Then, if we observe $X = 0$ or $X = n$, then the posterior distribution is improper.

This is particularly important in modern Bayesian models where high dimensional integration is often carried out using Gibbs sampling or MCMC.

Example 44

Consider the simple random effects model

$$\begin{aligned} y_{ij} &= \beta + \mu_i + \epsilon_{ij} \quad \text{where} \\ \epsilon_{ij}|\phi_\epsilon &\sim \mathcal{N}\left(0, \frac{1}{\phi_\epsilon}\right) \end{aligned}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, n_i$ where $\sum_{i=1}^{k} n_i = n$, and suppose that we use the improper priors

$$\begin{aligned} p(\beta) &\propto 1, \qquad & \boldsymbol{\mu} &\sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\phi_\mu}\mathbf{I}\right) \\ p(\phi_\epsilon) &\propto \frac{1}{\phi_\epsilon}, \qquad & p(\phi_\mu) &\propto \frac{1}{\phi_\mu}. \end{aligned}$$

Then, given the sample data, $\mathbf{y}$, we can show that the conditional posterior distributions are

$$\begin{aligned}
\beta|\mathbf{y},\boldsymbol{\mu},\phi_\epsilon,\phi_\mu &\sim \mathcal{N}\left(\bar{y}-\frac{1}{n}\sum_{i=1}^{k}n_i\mu_i,\frac{1}{n\phi_\epsilon}\right)\\
\boldsymbol{\mu}|\mathbf{y},\beta,\phi_\epsilon,\phi_\mu &\sim \mathcal{N}\left(\begin{pmatrix}\frac{n_1\phi_\epsilon(\bar{y}_1-\beta)}{n_1\phi_\epsilon+\phi_\mu}\\ \vdots \\ \frac{n_k\phi_\epsilon(\bar{y}_k-\beta)}{n_k\phi_\epsilon+\phi_\mu}\end{pmatrix},\begin{pmatrix}\frac{1}{n_1\phi_\epsilon+\phi_\mu} & 0 & \dots & 0\\ \ddots & \vdots & \vdots & \vdots\\ 0 & \dots & 0 & \frac{1}{n_k\phi_\epsilon+\phi_\mu}\end{pmatrix}\right)\\
\phi_\epsilon|\mathbf{y},\beta,\boldsymbol{\mu},\phi_\mu &\sim \mathcal{G}\left(\frac{n}{2},\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\beta-\mu_i)^2}{2}\right)\\
\phi_\mu|\mathbf{y},\beta,\boldsymbol{\mu},\phi_\epsilon &\sim \mathcal{G}\left(\frac{k}{2},\frac{\sum_{i=1}^{k}\mu_i^2}{2}\right)
\end{aligned}$$

where $\bar{y}=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}y_{ij}$ and $\bar{y}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}y_{ij}$.

Now, a Gibbs sampler could be set up by sampling sequentially from the various conditional distributions. However, the results do not make sense as the joint posterior distribution in this case can be shown to be improper. See e.g. Hill (1965).

There are a number of published papers using MCMC methods where the results are, in reality, meaningless as the posterior distributions are really improper.

The use of Jeffreys priors does not satisfy the likelihood principle.

Example 45

Suppose that we are going to generate binomial data $X|\theta \sim \mathcal{BI}(n, \theta)$. Then, from Example 36 we know that the Jeffreys prior is $\theta \sim \mathcal{B}(1/2, 1/2)$. Now suppose that we change the experimental design so that we will now generate negative binomial data. Therefore:

$$\begin{aligned}
\log f(X|\theta) &= c + r \log \theta + X \log(1-\theta) \\
\frac{\partial \log f(X|\theta)}{\partial \theta} &= \frac{r}{\theta} - \frac{X}{1-\theta} \\
\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} &= -\frac{r}{\theta^2} - \frac{X}{(1-\theta)^2} \\
-E\left[\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\right] &= \frac{r}{\theta^2} + \frac{r}{\theta(1-\theta)} = \frac{r}{\theta^2(1-\theta)}
\end{aligned}$$

The Jeffreys prior is

$$p(\theta) \propto \frac{1}{\theta(1-\theta)^{1/2}}.$$

Thus, Jeffreys prior depends on the experimental design and, if we observe 9 heads in 12 tosses of the coin, we need to know the design before the posterior distribution can be calculated. The posterior is $\theta|\mathbf{x} \sim \mathcal{B}(9.5, 3.5)$ given binomial data and $\mathcal{B}(9, 3.5)$ given negative binomial data.

Other problems

- Inadmissibility: Bayesian inference (with proper priors) leads to admissible estimators, but the use of improper priors can lead to inadmissible estimators.

  Example 46
  Suppose that $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{I}\right)$ and that we use a uniform prior $p(\boldsymbol{\theta}) \propto 1$. Then given an observation, $\mathbf{x}$, the posterior is $\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I})$ and the posterior mean, $\mathbf{x}$ is an inadmissable estimate of $\boldsymbol{\theta}$ if the dimension of $\mathbf{X}$ is greater than 2.

- Marginalization paradoxes (Dawid et al 1973), strong inconsistency and incoherence (Stone and Dawid 1972, Stone 1982).

References

Bayes, T.R. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53**, 370–418.

Berger, J. (2006). The case for objective Bayesian analysis (with discussion). *Bayesian Analysis*, **1**, 385482.

Brier, G.W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Buehler, R. (1971). Measuring information and uncertainty. In Godambe, V. and Sprott, D. eds. *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston.

Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society Series B*, **41**, 113–147.

Box, G.E. and Tiao, G.C. (1973). *Bayesian inference and statistical analysis*. Reading, MA: Addison-Wesley.

Chaloner, K. and Duncan, G.T. (1983). Assessment of a beta distribution: PM elicitation. *The Statistician*, **32**, 174-180.

Cooke, R.M. (1991). *Experts in Uncertainty*. New York: Oxford University Press.

Cooke, R.M., Mendel, M. and Thijs, W. (1988). Calibration and information in expert resolution: a classical approach. *Automatica*, **24**, 87–94.

Dawid, P., Stone, M. and Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference (with discussion). *Journal of the Royal Statistical Society, Series B*, **35**, 189–233.

DeGroot, M. and Fienberg, S. (1982). Assessing probability assessors: calibration and refinement. *Statistical Decision Theory and Related Topics, III*, eds. Gupta, S. and Berger, J. New York: Academic Press.

Finetti, R. de (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de lInstitut Henri Poincaré*, **7**, 1-68.

Fox, B.L. (1966). A Bayesian approach to reliability assessment. Memorandum **RM-5084-NASA**, The Rand Corporation, Santa Monica, USA.

Garthwaite, P.H., Kadane, J.B. and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, **100**, 680–701.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice (with discussion). *Bayesian Analysis*, **1**, 403420.

Hill, J.M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association*, **60**, 806–825.

Hughes, G. and Madden, L.V. (2002). Some methods for eliciting expert knowledge of plant disease epidemics and their application in cluster sampling for disease incidence. *Crop Protection*, **21**, 203-215.

Kahneman, D. Slovic, P. and Tversky, A. (1982). *Judgment under uncertainty: heuristics*

and biases. Cambridge: University Press.

Jaynes, E.T. (1968). Prior probabilities. *IEEE Trans. on Systems Science and Cybernetics*, **4**, 227–241.

Jaynes, E.T. (1983). *Papers on Probability, Statistics and Statistical Physics*. Ed. R.D. Rosenkrantz. Dordrecht: Reidel.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society*, **A. 186**, 453-.461.

Laplace, P.S. (1812). Teoría analítica de las probabilidades.

Lichtenstein, S., Fischoff, B. and Phillips, L. (1982). Calibration of probabilities: the state of the art to 1980. In Kahneman, D., Slovic, P. and Tversky, A., eds. *Judgement under uncertainty: heuristics and biases*. Cambridge: University Press.

Kass, R.E. and Wassermann (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343–1370.

Matheson, J. and Winkler, R.L. (1976). Scoring rules for continuous probability distributions. *Management Science*, **22**, 1087–1096.

Murphy, A. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595–600.

Price, P.C. (1998). Effects of a relative-frequency elicitation question on likelihood judgment accuracy: The case of external correspondence. *Organizational Behavior and Human Decision Processes*, **76**, 277-297.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Boston: Harvard University Press.

Russo, J.E. and Shoemaker, P.J.H. (1989). *Decision traps*. New York: Simon and Schuster.

Savage, L.J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, **66**, 781–801.

Stone, M. (1982), Review and analysis of some inconsistencies related to improper priors and finite additivity. *Logic, Methodology, and Philosophy of Science VI: Proceedings of the Sixth International Congress*, Hannover 1979, 413–426, Amsterdam: North-Holland.

Stone, M. and Dawid, A.P. (1972), Un-Bayesian implications of improper Bayes inference in routine statistical problems, *Biometrika*, **59**, 369–375.

Tversky, A. and Kahneman, D. (1973). On the psychology of prediction. *Psychological Review*, **80**, 237–51.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, **185**, 1124–1130.

Tversky, A. and Kahneman, D. (1980). Causal Schemas in Judgment Under Uncertainty. In Fischbein, M. ed. *Progress in Social Psychology*, Hillsdale, NJ: Lawrence Erlbaum, 49–72.

Wallsten, T.S., Budescu, D.V., and Zwick, R. (1993). Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science*, **39**,

176–190.

Winkler, R.L. (1986). On good probability appraisers. In Goel, P. and Zellner, A. eds. *Bayesian Inference and Decision Techniques*. New York: Elsevier.

Wiper, M.P. (1987). *The expert problem*. M.Sc. Dissertation. Department of Statistics, Manchester University.

Wiper, M.P. (1990). *Calibration and use of expert probability judgements*. Ph.D. Thesis. Department of Computer Studies, Leeds University.

Wiper, M.P., French, S. and Cooke, R. (1994). Hypothesis based calibration scores. *The Statistician*, **43**, 231–236.

Witteman, C. and Renooij, S. (2003). Evaluation of a verbalnumerical probability scale. *International Journal of Approximate Reasoning*, **33**, 117-131.

Yang, R. and Berger, J.O. (1997). A catalogue of noninformative priors. *Working Paper*, **97-42**, Institute of Statistics and Decision Sciences, Duke University. Available from: .