A primer in Bayesian Inference

Aart F. de Vos
draft September 2000, revision Februari 2008

1.1 Introduction

One of the most intriguing fundamental controversies in modern science is that between Classical and Bayesian Statistics. The controversy is of a philosophical nature. In Classical statistics truth is fixed and observations are random. Bayesian statements are probability statements about possible states of the truth. Bayes' formula states how to revise probability statements using data. And that is about all, it is amazingly simple. About Bayes' rule there is no disagreement. The rule is

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{1.1}$$

and follows directly from the definition of conditional probability: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$.

It works as follows:

Suppose a die is thrown under a dice-box. According to the standard model, all outcomes have probability 1/6. Now the dice-box is lifted a bit, and a (random) corner of the upper side becomes visible; it contains a dot. What is the new probability distribution of the outcomes?

A is the outcome of the throw, $A_i = 1....6; P(A_i\ )$=1/6.

B is: a randomly chosen corner contains a dot.

We obtain the following table:

A	$P(A)$	$P(B\|A)$	$P(A \cap B)$	$P(A\|B)$
1	1/6	0	0	0
2	1/6	1/2	1/12	1/8
3	1/6	1/2	1/12	1/8
4	1/6	1	1/6	1/4
5	1/6	1	1/6	1/4
6	1/6	1	1/6	1/4

The easiest way to construct the last column is to multiply. for each value of A, $P(A)$ and $P(B|A)$, to sum these values and divide by this sum. This last operation is called scaling, and corresponds to the formula as

$$\sum_A P(A)P(B|A) = \sum_A P(A \cap B) = P(B)$$

An easier argument is that $P(A|B)$ has to be a probability distribution, so sum to unity. As the scaling operation is trivial, Bayes rule is also written as

$$P(A|B) \propto P(A)P(B|A)$$

The symbol $\propto$ means "is proportional to"

The probabilities and probability distributions in this expression have names:

- $P(A)$ is the **prior (distribution)**: what is known about A before B is observed.
- $P(B|A)$ is the **likelihood.** Note that it only refers to the observed fact B, for all values of A. It is not a distribution!
- $P(A|B)$ is the **posterior (distribution):** what is known about A after observing B.

And the central theorem of Bayesians Statistics is that Statistical inference may be based on the simple device

$$posterior \propto prior * likelihood$$

In the example of die-throwing this is not of controversial. The discussions concern the possibility of using Bayes' rule as:

$$P(Truth \mid Data) = \frac{P(Truth)P(Data \mid Truth)}{P(Data)} \tag{1.2}$$

which tells you how to do inference the Bayesian way. You must be prepared to assign probabilities to "Truth", before having seen "Data", in other words you must specify

$$P(Truth) = the\ \text{“}prior\text{”}. \tag{1.3}$$

The second ingredient you need is data, plus an idea of how the data relate to the truth, which is nothing but the classical idea of specifying a stochastic relationship

$$P(Data \mid Truth) = the\ \text{“}likelihood\text{”} \tag{1.4}$$

for all relevant values of "Truth". Note that $P(Data \mid Truth)$ is not used as the probability distribution for different Data, but as the probability of the given data for different values of "Truth". Some authors do use $L(Data \mid Truth)$ for the likelihood to avoid this misunderstanding.

Now, noting that (using T for $Truth$), $P(Data)$ can be written as

$$P(Data) = \int P(T)P(Data \mid T)dT \tag{1.5}$$

that is as a function of $P(T)$ and $P(data|T)$, it is clear that the prior and the likelihood enable you, using (1.1) to construct

$$P(Truth \mid Data) = the\ "posterior", \tag{1.6}$$

a new probability statement about T given the data.

Bayesian inference thus shows how to learn from data about an uncertain state of the world (="truth") from data. And inference simply follows the laws of probability calculus.

All this may seem perfectly natural, but classical statistical inference is different:: there probabilities are only specified for $P(Data \mid Truth)$, and inference has concentrated on rules based on data that might arise for different states of $Truth$, represented in most cases by parameters.

1.2 A simple test revealing a shocking difference

We will see that many results of Classical Statistics have a similar Bayesian counterpart. But some inferences are strikingly different, specifically in testing.

A simple example is the famous test whether somebody can taste a difference between two drinks. Sir Ronald Fisher, the famous statistician, used the example of his aunt who was said to be able to taste whether the sugar had been put first into her tea and next the milk or vice versa. After blindfolding the aunt and randomizing marked cups of tea, the test is constructed from the distribution of the number of right guesses s (the data), given the hypothesis that the probability of a right guess, p, is just 1/2 (a possible truth).

The classical test uses the fact that if p=1/2, and n (the number of cups) is 20,

$P[s \geq 15|p = 1/2] = 0.02$

So the null hypothesis is rejected at the standard 5% significance level.

The Bayesian test has to be more specific about the alternative: the probability of a right guess has to be specified. Say p=0.75 is the alternative (one may also specify a prior distribution for p in the alternative case, but that requires some more calculations).

The most elegant way to calculate the posterior probabilities is **Bayes' rule for two alternatives:**

$$\frac{P[T|D]}{P[-T|D]} = \frac{P[T]}{P[-T]} . \frac{P[D|T]}{P[D|-T]} \tag{1.7}$$

or in words:

$$\textbf{posterior odds} \ = \ \textbf{prior odds} \times \textbf{likelihood ratio}.$$

where the likelihood ratio is also called the "**Bayes Factor**".
In the example, D is " s=15", T is "p=0.5", and $-T$ is "p=0.75"
The Bayesian posterior odds then follow from

$$\begin{aligned} \frac{P[p = 0.5|s = 15]}{P[p = 0.75|s = 15]} &= \frac{P[p = 0.5]}{P[p = 0.75]} \frac{P[s = 15|p = 0.5]}{P[s = 15|p = 0.75]} \\ &\cong \frac{P[p = 0.5]}{P[p = 0.75]} . \frac{1}{13.7} \end{aligned}$$

The final answer depends on the subjective prior odds. Prior odds of 99:1 (a 1% probability that the aunt has the ability) change by the result to posterior odds of 99:13.7, a 12% probability. This is quite different from the 95% (or 98%, the "p-value") one might think that the classical answer means.

Also for a prior 1:1, suited for situations where one has no idea, the posterior odds are 13.7:1, so 6.7%, still rather different. But of course the alternative of p=0.75 is rather high.

A fundamental difference between the two approaches is that classical statistics concentrates on the full set of data that might arise, while Bayesians concentrate on the observed data only. This is called the "**Likelihood principle**": **inference should only refer to the data actually observed.** So not to "15 *or more*". For the same reason the fact that 0.75 (=15/20) is an "unbiased" estimate of p, is in itself irrelevant in Bayesian eyes: it refers to repeated sampling under p=0.75.

A related difference is that Bayesians may continue sampling until they wish to stop, and then simply use the n and s at that moment for their conclusion (this is the "**Stopping Rule Principle**": data based stopping rules are irrelevant for inference about a parameter; this is a major difference with classical results). Moreover the Bayesian answer can be built up in two steps: after observing s_1 successes in the first trial, the resulting posterior can be used as prior in the next experiment:

$$\begin{aligned} P(p|s_1,s_2) &\propto \pi(p)P(s_1,s_2|p) \\ &= \pi(p)P(s_1|p)P(s_2|p) \\ &\propto P(p|s_1)P(s_2|p) \end{aligned}$$

so sequential inference has a very nice form. Note however that in this case $P(s_1,s_2|p) = P(s_1|p)P(s_2|p)$:the outcomes are conditionally (given p) independent. In general $P(s_1,s_2|p) = P(s_1|p)P(s_2|s_1,p)$, which is more complicated.*Exercise 1:*

A. check the calculations in the "cups of tea" example with n=20, s=15 and compute outcoms for the alternatives p=0.6 and/or s=14.

B. Compute in a spreadsheet all combinations of prior odds and values of p under the alternative such that the posterior probability that "aunt can do it" is .02.

C. Suppose the experiment was: start with 10 cups of tea, if the p-value is between 5% and 30% then take a second sample of 10. What is then the "error of the first kind? And what changes in the Bayesian answer?

D. Suppose sampling has continued until the result was significant. The Bayesian answer does not change! How to explain this to a classical statistician?

1.2.1 More serious: law cases.

In England, several mothers have been convicted to a life sentence because two of there children had died, possibly by murder. The other explanation was a double case of the "Sudden Infant Death Syndrome (SIDS)". The basis of the verdict was that twice SIDS is extremely unlikely. So, the null hypothesis (innocence) was rejected. The prior odds (how likely is it that a mother murders her children were not considered. In 2002, one of the cases, "Sally Clark" was reopened, and the prominent Bayesian statistician Phil David (see his website for the article "weighing evidence by juries") convinced the judge of the erroneous reasoning. Sally Clark, and a number of others were released. In the Netherlands I started a similar discussion in 2004 about the case "Lucy de B", a nurse with a suspect high number of death cases while she was on duty. In vain, perhaps justified as there was more evidence than that. Her life sentence was confirmed in the appeal case. Most striking was the resistence of Dutch statisticians against Bayesian reasoning.

1.3 Some history

Probability statements about the truth are the natural context for decision making under uncertainty. If one can formulate utilities for all combinations of decisions and the state of nature, one can optimize expected utility. Though there is much discussion about details, this schedule is the dominant rational basis for decision making under uncertainty. Leonard Savage, in "the foundations of statistics" (1953) provided the axiomatic underpinning.

With respect to statistics, it appears that estimation of parameters, confidence intervals and hypothesis testing may all be formulated as solutions to decision problems along Bayesian lines. In a way Bayesian analysis is much simpler than classical analysis: the same approach is used anywhere. However, at the cost of specifying priors.

"Classical" statisticians have problems with the idea that a prior probability statement on "Truth" is needed. In their view Truth is nonstochastic, and they try to define procedures with good properties for *any* Truth. In Bayesian eyes, this is impossible in general, and the "false idol of objectivity" (Leamer) has led to many misunderstandings.

Bayesians postulate priors as existing, subjective "degrees of belief". In some statistical problem these priors matter much, sometimes (when there are many data) the priors are practically irrelevant. In the latter case, classical and Bayesian results are similar. This similarity may also be attained in another way. Specifying priors that are "noninformative" lead, especially in estimation problems, to remarkable dual results with classical outcomes. Thus most classical statistical procedures get a new, Bayesian interpretation. With this interpretation, many problems in classical statistics disappear. The basis of this school is led bay Sir Harold Jeffreys (1939), the work by Box & Tiao (1973) was the breakthrough to a larger public.

Since, an ever increasing number of scientists came to the conviction that statisticians should abolish their classical paradigms, and replace it by Bayes' theorem. However, despite the obvious successes of the Bayesian approach, including conversion of many outstanding statisticians to the Bayesian "religion",

mainstream statistics is classical. Statistical education starts with the classical line, Bayesian inference remains something special, mistrusted by many for its subjective character.

The paradigm debate has its roots in papers by Thomas Bayes (an English reverend), published posthumously in 1763, and Laplace, independently written in 1775. But until Jeffreys' (1939) work, debate was muted. In the 50's and 60's work by i.a. Savage and Lindley propagated the Bayes approach, and after Box and Tiao(1973) the paradigm debate became quite general, loosing little of its sharpness. Practical, philosophical and feasibility considerations all play a role. For a survey of the history, see Stigler(1997). For sharp discussions see e.g. Berger(1986) propagating "Bayesian salesmanship" and Efron(1985) explaining "why isn't everyone a Bayesian". Beautiful survey articles, stressing the fact that Bayes is a paradigm -and the only good one -are Lindley (1990) and Bernardo (2003, free download from his homepage). A readable, complete and up to date book is O'Hagan (1994). More advanced and concerned with foundations is Bernardo and Smith (1994).

In **econometrics**, Zellner(1971) for a long time was the only textbook based on Bayesian principles. "Specification Searches" by Leamer(1978) is a more advanced classic in Bayesian econometrics. Poirier(1995) is a nice book about statistics, with some references to econometrics. Lancaster(2004) is a very accessable book for econometricians. Geweke(2005) is more advanced with a lot of attention for computational issues.

In journals, Bayesians have long been a (suppressed?) minority. Poirier ("a report from the battle front", 1989) analyzed statistical and econometric journals in the period 1982-1986. Only 7.8 percent of the pages were devoted to Bayesian articles. But the percentage keeps growing, by 2004 25% of the article at least mentioned Bayes (source Poirier, conference paper).

The major breakthroug after 1990 was **computational**. Anaysing models with many parameters was a huge problem, until the possibilities of **MCMC** (Markov Chain Monte Carlo) simulation were discovered. The Gibbs Sampler and the Metropolis-Hastings algorithm make it possible to analyse huge models and as computers get ever faster, the sky is the limit. The possibilities have brought many non- or semi-Bayesian statisticians and econometricians (e.g. Niel Shephard, who invented some important algorithms for time series) to use it. Especially for models with latent (not directly observed) variables MCMC methods are ideal as estimates of these latent variables are obtained.

The best known **computer package** is **BUGS** (Bayesian analysis Using the Gibbs Sampler), developed by David Spiegelhalter. See the website. A well documented free version of WINBUGS, suited for not too big models can be downloaded. The orientation was primarily medical statistics, but many econometric applications are possible now (see Lancaster(2004)).

Still, model building outside the context of decision making is less convincing than the study of real problems where much uncertainty is involved and decisions still have to be made (law, auditing etc). There subjective elements necessarily play an important role. In these situations **Bayesian Decision Analysis** provides the natural way to process prior and statistical information from different sources. The posterior gives the probabilities of all possible states of nature given the information used. By calculating the results of decisions in each of the possible states, and by assigning "utilities" to all possible outcomes (a form of cost-benefit analysis) one arrives at a framework for optimal decisions. This framework, the maximization of "expected utility", is for Bayesians the only coherent way to make decisions. It is the basis of "Statistical Decision Theory", having its roots in a seminal book by Raiffa and Schlaifer (1961), updated (and simplified) in Pratt, Raiffa and Schlaifer (1995). Unfortunately, this theory has developed as a more or less separate branch of science.

To get the real flavour of Bayesian thinking one should study a field like auditing. In auditing a standard situation is that no or little errors are found and the main question is how large a sample must be to have good confidence that the population contains little errors, but the questions "how likely is it that this population contains many errors given that no errors are found?" and "how likely is it to find no errors if there are many errors?" are systematically confused (while the answers differ strongly). For solutions of the audit problem, defined as optimal sampling from cost-benefit perspective see Wille(2003).

Confusion about what questions are answered in statistical inference is quite wide spread. This has done much harm, I think. To be successful in a business environment, econometricians must learn that their data series contain only a part of the necessary information needed for decisions. They must learn how to cope with model uncertainty. They must not rely too much on asymptotic properties. They must be able to give an honest picture of forecast uncertainty. They must be able to confront their ideas with the prior knowledge of experts. And above all, they must have a clear picture of the road that leads from models to decisions. All this may be learned by studying Bayesian inference.

1.4 Bayes' rule and subjective probability

Bayesian Statistics departs from the view that we are uncertain about the true state of affairs, and that this may be expressed by probability statements

about the truth. The revision of these probability statements using data is the uncontested domain of Bayes' rule. The only problem is that one has to start with a probability statement before any data are available: the prior distribution. This is the central subjective element.

We begin with a simple example of Bayesian Inference, describing the most relevant aspects and specifically how it points the way in acquiring information to get empirically justified priors. The question we adress is:

Is it cold outside?

A. Prior beliefs and making bets.

Suppose you were just outside and you judge that the temperature was about zero. You are rather certain that it was between -2 and +2 degrees. Say 95% sure. What does this statement mean? A plausible interpretation would be that you are prepared to make a bet: I pay you \$1 in the event that the temperature is between -2 and +2 and you pay me \$19 if this is not true: This implies "betting odds" of 1:19" . If you are not prepared to accept less favorable bets, you apparently think that the probability that the temperature is outside the specified interval is 1 in 20, which makes your expected profit $\$1 \cdot (19/20) - \$19 \cdot (1/20) = 0$. This definition of subjective probabilities in terms of implied decisions is primitive, but illustrates the idea. At the end of this section we will discuss this in more detail. For the moment it is sufficient to argue that it is possible to specify a prior. If we are prepared to assign a probability to the event that the temperature is between -2 and +2 degrees we may as well assign probabilities to other intervals as well. In doing so, we will end up with a complete probability distribution, for instance a normal $\mathcal{N}(0,1)$ distribution. No fundamentally new steps are involved in this extension of the notion of subjective probability.

It is important to realize that the probability statements we are making reflect the information they are based upon. If you had not been outside for some time, it would have been much harder to bet. If you were outside some hours ago a subjective probability distribution $\mathcal{N}(0,2)$ might be more appropriate. On the other hand, if you saw a thermometer hanging outside indicating about 0.5 degrees your probability statement would be $\mathcal{N}(0.5,0.1)$.

Still, by the way, a probability statement: for most events an exact true value may never be attained. But whether a true value exists or not is not essential: let somebody guess your age in years -established beyond doubts I suppose- and (s)he will come up with a probability statement (the next step is telling when your mother was born and asking how this information must be used, after which you almost certainly have to explain Bayes' rule)

Exercise 2: Give a probability statement about your teacher's age, and process the further information he will give to you.

B. New information and Bayes' rule.

Back to the situation where your prior was $\mathcal{N}(0,1)$, based upon what you felt when being outside. Suppose now that you look out of the window and you see that it rains. To simplify things let's assume that rain only occurs when the temperature is above zero and not when it is below zero. So obviously your observation tells you something about the temperature. The question is: what bets are you now prepared to make? Or in other words: what probability statement can you make now?
Here we have to use Bayes' formula. Defining T as the temperature and R as the event that it rains we get

$$P(T \mid R) = \frac{P(R \mid T)P(T)}{P(R)} \tag{1.8}$$

Here $P(T)$ is the prior distribution: what we could say about the temperature before we observed that it rained. $P(T \mid R)$ is the posterior distribution: what we may say about T after having observed that it rains. What we need to process is the information $P(R \mid T)$, the probability of rain for different temperatures. $P(R)$ is not needed, it follows from $P(T)$ and $P(R \mid T)$ by

$$P(R) = \int P(R \mid T)P(T)dT \tag{1.9}$$

We may make the following setup (using instead of the normal distribution a rough discrete approximation, and an educated guess $P(R \mid T) = 0.1$ for $T > 0$):

$T(emp) \in$	$(-3,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,3)$
$Prior$	0.025	0.100	0.375	0.375	0.100	0.025
$P(R \mid T)$	0	0	0	0.1	0.1	0.1
$P(T \mid R)$	0	0	0	0.750	0.2	0.050

To calculate $P(T \mid R)$ an important trick is used. We just multiplied $P(T)$ and $P(R \mid T)$ for each value of T. This gives $P(T \mid R)P(R)$. As $P(T \mid R)$

defines a probability distribution for T, and $P(R)$ does not depend on T, we may just scale the outcomes: divide them by their sum, such that they sum to unity. (Note that this sum is just $P(R)$, 0.05 in this case). This trick is also used to shorten the notation of Bayes' formula to:

$$P(T \mid R) \ \propto \ P(T)P(R \mid T). \tag{1.10}$$

The posterior $P(T \mid R)$ shows that you should be prepared now to bet that $P(T > 2)$ at 1:19 compared to 1:39 in the previous case. The example has an interesting special feature. If we had used values for $P(R \mid T > 0)$ different from .1, but equal for all $T > 0$, we would have obtained the same $P(T \mid R)$. The only relevant assumption appears to be that the probability of rain does not differ for different $T > 0$. The lesson is that information only changes knowledge if the observed fenomenon (R) has different probabilities ($P(R \mid T)$ for different values of the object of knowledge (T). If rain is equally likely for all relevant values of T, it is irrelevant information. "Equally likely" is, by the way, for Bayesians a typical way to express a lack of prior knowledge. Obviously it would in our example be fruitful if you knew more about the probability of rain fall at different temperatures. From a careful study of the weather statistics for similar circumstances you might learn:

$T(emp) \in$	$(-3,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,3)$
$P(R \mid T)$	0	0	0.01	0.11	0.09	0.08

Note that we are only interested in the intervals between -3 and +3: the search for relevant information may be guided by available prior information and by the goal of the analysis.

Exercise 3: Compute $P(T \mid R)$ with this new information. One should notice that the new $P(R \mid T)$ does not differ much from the old one. Try to formulate why and when this is the case in general.

C. Information in the form of an inequality.

The information that it rained without further reference to the probability of rainfall at different $T > 0$ was implemented by attaching equal probabilities to $P(R \mid T)$ for different T. The versatility of Bayesian inference is illustrated by the following alternative to processing the information that it rains:
Suppose we only know that the temperature is above zero when it rains, but nothing about $P(R \mid T)$ for different $T > 0$. It is possible to use this information directly:

We know that $T > 0$; let C be the event $T > 0$. Then

$$P(T \mid C) = \frac{P(C \mid T)P(T)}{P(C)} \tag{1.11}$$

$$P(C) = \int P(C \mid T)P(T)dT \tag{1.12}$$

which is nothing but the probability that $T > 0$ in the prior. Clearly (1.11) is simply the truncated prior of T. This is an important result, as inequality restrictions occur frequently in statistical practice and tends to cause problems in the classical setup (see Box and Tiao(1973) section 1.5). Noteworthy is that we get exactly the same result as we got assuming that $P(R \mid T)$ was 0.1 for $T > 0$. In a way this confirms the intuitive notion that being unable to be specific equals attaching equal probabilities to all possibilities.

D. Empirical underpinning of the prior.

Perhaps the most important lesson thus far is that Bayes' rule leads the way in information processing. Any aspect of the problem has to be put into probability statements. This can be a subjective statement but in many cases it is possible to search for information that replaces some of the subjective part by empirical information. This process is our main concern: the implementation of Bayesian methods in a way as closely as possible connected to real experiences. Let's investigate the temperature statements further.

The prior distribution for T in (1.8) was based upon a personal estimate of temperature which one felt. An empirical base might be given to this by systematic research: write down daily the assessed temperature feeling, followed by a detailed measurement. Thus one gets an idea of the bivariate distribution of subjective and true temperature. We call t the subjective point estimate and T the true temperature. From this we may obtain $P(T \mid t)$, the prior distribution in our example. For $t = 0$ may simply come from the distribution of T at all days we thought $t = 0$.
This may seem trivial but there is a snag. We must have $P(T \mid t)$ and not $P(t \mid T)$. A classical statistical approach would be concerned $P(t \mid T)$: as t is the subjective temperature and T the real one the only appropriate model seems to be $\underline{t} = T + \underline{u}$ with $\underline{\text{u}}$ u a disturbance term, let's say a $\mathcal{N}(0,\sigma^2)$ distributed. For the classical approach it is essential to discern between stochastic variables (underlined in the previous sentence) and nonstochastic variables (T,

being the true temperature). In the Bayesian approach this distinction is impossible: every variable is stochastic (so we use no underlining or capitals for stochastic variables).
A Bayesian would, believing that regularities could be discovered in $P(t|T)$, model this as well (to use more information). As the next step he would convert this into the relevant $P(T \mid t)$ by

$$P(T \mid t) = \frac{P(t \mid T)P(T)}{P(t)} \tag{1.13}$$

with

$$P(t) = \int P(t \mid T)P(T)dT \tag{1.14}$$

And remark that, beside a model for $P(t \mid T)$ a prior for $P(T)$ is needed: a probability statement on T based on information other than what we felt or whether it rains. One might for instance think of the distribution of the temperature based only on information about the time of the year, to be based empirically e.g. on the temperatures on the same day of the year in the last thirty years. Again, Bayesian inference appears to point the way in information processing.

E. Successive processing of information.

The processing of the information from T, the subjectively assessed temperature, as described above may be done in steps. It is illuminating to review these steps in a simplified discrete form. We start with the probability statement on T from experiences in the last years.

$T(emp) \in$	$(-\infty,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,\infty)$
$P(T)$	0.40	0.02	0.03	0.03	0.02	0.50

Seems a reasonable outcome. The next problem is to incorporate the model $P(t \mid T)$. To simplify things we assume that in forty percent of the cases the guess is right, in 40 percent one degree too low and in 20 percent one degree too high, independent of the true temperature. Suppose $0 < t < 1$ is 'observed' the relevant information is then:

$T(emp) \in$	$(-\infty,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,\infty)$
$P(t\|T)$	0	0	0.20	0.40	0.40	0

and we obtain $P(T \mid t)$ by multiplication of $P(t \mid T)$ with $P(T)$ followed by the scaling operation: sum all these products and divide them all by this sum:

$T(emp) \in$	$(-\infty,-2)$	$(-2,-1)$	$(-1,0)$	$(0,1)$	$(1,2)$	$(2,\infty)$
$P(T)$	0.40	0.02	0.03	0.03	0,02	0.50
$P(t\|T)$	0	0	0.20	0.40	0.40	0
$P(T\|t)$	0	0	0.23	0.46	0.31	0

and this posterior $P(T|t)$ is the prior for the next step where the information on the rainfall is incorporated like before.
If finally somebody comes with a thermometer and measures accurately the temperature outside, this information gives a likelihood of one in the interval where the measurement lies, and zero elsewhere (or better: almost zero, a thermometer might be defect). And so looks the posterior, the evidence "dominates" the prior.
In this way, different sources of information are processed sequentially. As might be expected it does not matter which information is used first: the final outcome is the scaled product of a series of information outcomes. The posterior after one round of using information is the prior for the next round. And this process may also be followed backwards: one may try to reformulate a prior as a posterior of some information process, thereby reducing the role of subjectivity.
There is one snag in this successive use of information: no information must be used twice. If you heard the weather forecast this morning and you just felt the temperature outside, your opinion is already influenced by the forecast. What you would like to know for a careful analysis is the opinion you would have had not knowing the weather report, to combine this with the information from the report. But this may be very difficult. Unfortunately there is no way to be certain that a prior does not depend on information to be used subsequently. Independence is an assumption, allowed if no dependency is plausible.

F. Decisions.

What is the use of all this apart from deciding whether to involve in betting? The simple answer is that we need probability statements about possible states of affairs to make rational decisions under uncertainty. The temperature case provides an example: suppose you have to choose whether to go by car or by train to work, and you are afraid of bad weather conditions. If the temperature is below zero, there is more risk of accidents by freezing rain.

The best scenario is going by car without freezing rain. We give this unit utility.

Going by car with freezing rain is the worst scenario, it gets utility zero.

Suppose that the probability of freezing rain (P(F)) for $-1 < T < 0$ is 10% (given that you observed rain). Then going by car with $-1 < T < 0$ gives expected utility 0.9. And if $P(F|0 < T < 1) = .01$, utility of going by car for this temperature is 0.99. We assume further $P(F|1 < T < 2) = 0$: going by car then gives unit utility. But we are uncertain about T , we only have a probability distribution of T. Going by car gives probability of freezing rain $\sum_i P(F|T_i)P(T_i)$ with T_i the three possible states of the temperature. The result, given $P(T)$, (see table and check) is $P(F) = 0.0276$. The construction of utility is such that, going by car, expected utility is $1 - P(F) = 0.9724$. Going by train, freezing rain is no problem. To make decisions, one must specify an equivalence relation, e.g. "going by train is equivalent to going by car with 2% chance on freezing rain". So expected utility (U) is 0.98, and one should go by train. The table in terms of expected utilities is:

	$T \in (-1, 0)$	$T \in (0, 1)$	$T \in (1, 2)$	
$P(T)$	0.23	0.46	0.31	
U|By car	0.9	0.99	1	$E(U) = 0.9724$
U|By train	0.98	0.98	0.98	$E(U) = 0.98$

This way of constructing utilities is a theoretically famous one. All possible outcomes are made equivalent to a choice between heaven and hell (utilities 1 and 0). If this equivalence exists, a composite outcome, like going by car with different possible outcomes for the temperature, is also equivalent to such a choice, with "expected utility" as chance on "heaven". This formulation makes it obvious that one should maximize expected utility.

G. More on utility.

The formulation of utility in terms of equivalence between one option with certainty and a probabilistic choice between two extremes is too theoretical for practice. In practice other utilities, often in money terms are used. As any linear transformation on the utility scale ($u' = a + bu, b > 0$), does not matter for the resulting decisions, one may choose a convenient representation.
In economics, profit maximization is so strong a paradigm that the maximization of expected utility is often confused with maximizing expected profits. This is not a good thing to do in general. Why do people insure themselves against financial disasters for premiums that imply an expected loss for them (the profit of the insurance company)? Because a loss of $10,000 with probability 1/1000 is worse than a loss of $10 with certainty.
The "utilities" of the resulting situations are not proportional to the amount of money possessed. Some say one should use its logarithm as utility, anyhow the utility of $1 extra decreases the more dollars you own. On the other hand for the insurance company the utility of one insurance is almost proportional to the expected profit. The difference in utility evaluation between client and insurance company explains the important role the latter plays in society.

H. The foundations of probability.

We started our example with identifying probability statements with bets, we ended with a structure for probability, utility and decisions. In between it became obvious that probability statements about the same phenomenon change with different information sets. Obviously we use other notions of probability than the classical "long run frequency", or the mathematical postulate that "probabilities are entities satisfying the axioms of probability calculus". The search for other foundations has inspired many scientists, and by now about fifteen different foundations have been formulated. Reny(1975) formulated axioms containing the idea that probabilities always depend on information. Savage (1962) formulated requirements for decisions, leading to the definitions of utility and probability.

That (partly) subjective probabilities have to fulfill the axioms of probability calculus, is called the *principle of coherence,* first formulated by de-Finetti (1935), based on the betting concept (see for alternatives Bernardo and Smith(1994)). The principle says that probability statements should correspond with the willingness to engage in all resulting bets, and that it must be impossible for another party to construct a series of bets such that he wins for any realisation.

Such a bet that allways wins is called a "Dutch book". A nice example of a Dutch book occurs when a betting agency offers bets on "what country will be world champion football?": if in six countries people bet 1:4 on their own country, the bookmakers take six bets, recieve 6 and pay 4 whatever happens.

The duality between bets and probabilities can be established in several ways. A bet is defined as paying a stake $S(A)$ to get 1 if A happens. Your personal probabilities determine whether you are prepared to bet; for $S(A) = P(A)$ your expected profit is zero. It is assumed that from your behaviour in betting your $P(A)$ may be deduced, e.g. by taking for $S(A) = P(A)$ the maximal stake you are prepared to pay betting on A.

The coherence principle says that these stakes have to fulfill the axioms of probability calculus:

The three axioms for (finite) probabilities are:

If A and B are events, and U is the certain event, then

$$0 < P(A) < 1;\ P(U) = 1;\ P(A \cup B) = P(A) + P(B)\ if\ A \cap B = \oslash \tag{1.15}$$

Corresponding stakes have these properties, or else a dutch book is possible:

$S(A) > 0$ is trivial: if S(A)<0, your opponent bets on A, gets $-S(A)$ and 1 more if A happens

$S(A) < 1$ likewise: you must bet on A and if you pay $S(A) > 1$ you always loose.

$S(U) = 1$ if U always happens is also simple.

More complex is that if two events A and B are disjunct ($A \cap B = \oslash$), then $S(A \cup B) = S(A) + S(B)$.

If your opponent, confronted with your three stakes $S(A)$, $S(B)$ and $S(A \cup B)$, may involve in a combined bet with stake $c_1 S(A) + c_2\ S(B)\ + c_3 S(A \cup B)$, getting $c_1 + c_3$ if A happens, $c_2 + c_3$ if B happens and 0 if neither happens (c_i may be negative), he can choose the weights such that his profit is a if A happens, b if B happens and c if neither happens, by choosing a, b and c and solving:

$$\begin{bmatrix} 1 - S(A) & -S(B) & 1 - S(A \cup B) \\ -S(A) & 1 - S(B) & 1 - S(A \cup B) \\ -S(A) & -S(B) & -S(A \cup B) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Only if the matrix is singular, this is not possible, which is the case iff $S(A \cup B) = S(A) + S(B)$.

So the three axioms of probability calculus can be replaced by one principle.

Exercise 4: Prove in a similar way that coherence implies the definition of conditional probability $S(A|B) = S(A \cap B)/S(B)$. Note that is a conditional bet, if B does not occur, there is no bet at all. The relevant outcomes are $A \cap B, -A(not\ A) \cap B,\ and - B(not\ B)$.

1.5 The crucial step in modeling: making parameters stochastic.

Bayes' rule as given in (1.1) is in itself not a controversial issue, it follows directly from the definition of conditional probability. What matters is: what are we prepared to consider as being stochastic. The basis of Bayesian statistics is to make the object of analysis stochastic, and this object is generally something that in the classical approach is simply a "truth", as in (1.2). The temperature outside undoubtedly has some true value, but it is useful to work instead with the concept "our idea about the temperature outside", which is random and depends on the information we have.

A. Parametric models.

Usually the object of statistical inference is something less concrete than the temperature outside. In most cases "parameters" are the object of study. These parameters characterize a model that is assumed(!) to describe the possible outcomes of some phenomenon. In other words: not the phenomenon but its probability distribution is the object of inference.

Does this make much difference? Let p be the probability that the next toss of a coin results in "head". We may consider the toss or we may consider p. The toss -says classical statistics- is stochastic, p is not. But we may as well consider the fraction heads in the next billion tosses. This is still stochastic, while it is in the same time, apart from negligible deviations, equal to p. So there is no fundamental difference between outcomes of tosses and p. It is perfectly justified to assume a prior distribution for p. One may, defining this prior, think of the values of p that different coins possess. Obviously in the case of coin tossing this will be a prior concentrated around $p = 0.5$. Not however $p = 0.5$ with certainty : some, if not all, coins are "biased", albeit only marginally. In some way we must use our experience to state these degrees of belief. This may be done subjectively or empirically by studying, say, 1000 outcomes of 1000 different coins.

The Bayesian analysis proceeds by adding to the prior for p, denoted $\pi(p)$, the information from a series of experiments with the coin of which one wants to know p. This leads to a posterior distribution for p: if h heads resulted from n tosses we know

$$P(p \mid h) = \frac{\pi(p) P(h \mid p)}{P(h)} \propto p^h (1-p)^{n-h} \pi(p) \tag{1.16}$$

where the proportionality constant may be calculated afterwards, making $P(p \mid h)$ integrate to one.

Exercize 5: Take the following discrete prior for p:

p	0.47	0.48	0.49	0.50	0.51	0.52	0.53
$\pi(p)$	.0009	.009	.09	.8	.09	.009	.0009

Calculate in a spreadsheet $P(p \mid h)$ for $n = 100$, $h = 55$ and, $n = 1000$, $h = 550$; and for $n = 10,000, h = 5,500$. Make prictures containing the prior,

the scaled likelihood and the posterior

As is to be expected the result is a mixture of the prior ideas and the sample outcome. The larger the sample, the closer the posterior will be centered around the success rate in the sample; in this case the prior is very informative (concentrated around .5, so it requires many data to make the posterior different.

That this is a sensible setup becomes obvious if we wonder how we react with betting odds on the result of coin flipping: if 55 out of 100 is head, few people will deviate much from $p = 0.5$, but what to do if 90 out of 100 turns to be head?

In other situations with repeated successes and failures in equal circumstances we will react differently on experiences. Probability statements on the success rate of a medical operation will be rather vague as long as no operations are carried out. From similar events one will have some idea, but if 10 out of 15 operations succeeded the degree of belief in the success of the next operation will be close to 10/15. Clearly there is different prior information, giving a more important role to the information from the sample. Conclusion: the Bayesian setup conforms the way one feels sample information should be used.

B. Bayes' rule applied to parameters.

$$\begin{aligned} P(\theta \mid x) &= \pi(\theta)\frac{P(x \mid \theta)}{p(x)} \\ &= c\pi(\theta)P(x \mid \theta) \\ &\propto \pi(\theta)P(x \mid \theta) \end{aligned} \qquad \text{(bay)}$$

Or: posterior is proportional to prior times likelihood.

This simple formula opens wide perspectives. Moreover -with some generalizations - it is all one needs to solve any parametric statistical problem. Anyone who learned classical statistics with all its different approaches to the same problem can experience a miracle just by applying (1.17). The only problem is the prior. In some cases, unfortunately, this is a big problem.

C. Bayesian inference with normal prior and normal likelihood.

Suppose a sample $x_1 \dots x_n$ from a $\mathcal{N}(\mu\, \sigma^2)$ distribution, σ^2 known. And a prior for μ:

$$\pi(\mu) = \mathcal{N}(\mu_0, {\sigma_0}^2),$$

Then one may easily derive that the posterior is also normal.

Exercize 6: do this.

The nicest formulation of the result is in terms of **precisions.** (Precision, is the inverse of the variance). First

posterior precision = prior precision+data precision.

Or

$$\sigma_1^2 = Var(\mu|x) = \left(\frac{n}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)^{-1}$$

Second:

posterior expectation = λ***prior** exp**ectation** $+(1-\lambda)$***sample mean**

$$\textit{with } \lambda \textit{ the "relative precision"} : \ \lambda = \left(\frac{1}{{\sigma_0}^2}\right) \Big/ \left(\frac{n}{\sigma^2} + \frac{1}{{\sigma_0}^2}\right)$$

This formula has many interesting applications, examples are the Kalman Filter and random effect models.

D. Duality between Classical and Bayesian results.

The outcomes of the preceding sections have an important implication:

If n is large one may say that ${\sigma_0}^2$ is irrelevant, and one may as well delete it. This is an example of an important "law": *Asymptotically, Classical and Bayesian approaches are, except for interpretation, equivalent.*

Another possibility leading to the same result is putting the prior precesion zero (so σ_0^2 infinite) This approach is known as using a "noninformative prior". And this is an example of a second "law": *Some classical results have a dual Bayesian result when the prior is noninformative.* The dual Bayesian result in the Normal case is

$$\mu|x \cong \mathcal{N}(\overline{x}, \frac{\sigma^2}{n}) \qquad (1.17)$$

The duality is most clear in terms of confidence intervals: both the classical confidence interval and the Bayesian confidence interval are:

$$P\left(\mu \in \left[\overline{x} \pm z(\alpha)\sqrt{\frac{\sigma^2}{n}}\right]\right) = 1-\alpha \tag{1.18}$$

But the interpretation differs: in the classical case the non-stochastic μ lies between stochastic bounds, while in the Bayesian case the stochast μ lies between known bounds and what is more, as we know the full distribution of μ, we may take any interval we want. Moreover we are no longer bound to the prior choice of a confidence level of 95% or 99%, we simply may consider all levels we think important. So the Bayesian result is more useful. Moreover, the Bayesian interpretation seems more natural: $\overline{x}$ is just a figure, substituted into (1.18) it is strange to think of this figure as stochastic.

E. Survey of classical and Bayesian inference on parameters.

Statistical inference is concerned with statements about the unknown parameters. Once a model is assumed, the way classical statistics works may be described as follows:

- The existence of true parameters is assumed.
- The truth generates stochastic observations.
- Estimators or other statistics are functions of the stochastic observations.
- The properties of statistics thus depend on the true parameters and on sampling variation.
- The goal is to find statistics that lie close to the true values whatever these values are.
- In general this is not possible.
- Imposing restrictions (like unbiasedness), choosing a decent model (like one from the exponential family) and defining an expedient optimality criterion (like minimal variance) may reduce the problem to one that has a solution that is uniformly (i.e. independent from the true values) optimal.
- Some other solutions -in particular maximum likelihood solutions- may be proven to have approximately optimal solutions, in particular "asymptotically" which means that in large samples the solutions are close to optimal solutions. How large the samples must be and how close the solutions are however is generally unknown.

Bayesian Inference does not suppose true parameters. The starting point is a probability distribution of the parameters, the Prior Distribution. The data serve to change this idea of the truth into a new, generally more concrete idea: the Posterior Distribution. The formula that generates this transformation is Bayes formula:

$$P(\theta \mid x) = \pi(\theta)\frac{P(x \mid \theta)}{P(x)} \propto \pi(\theta)P(x \mid \theta) \tag{1.19}$$

Here $\pi(\theta)$ is the prior distribution: what we know about θ before we have seen the sample. $P(x \mid \theta)$ is the likelihood function. $P(\theta \mid x\ )$ is the posterior distribution: what we know about θ from the combined information of the prior and the sample. The proportionality constant follows from the fact that $P(\theta \mid x\ )$ must integrate to one.
Inference follows from the posterior in a coherent way. Confidence intervals may be derived directly from the posteriors (though one might wonder whether they are needed). The probability distribution of future values (section 1.7) is a straightforward result without classical parallel. Point estimation will appear to be the outcome of a decision problem (section 1.8), and hypothesis testing as well (section 1.9).

F. The Regression model with multivariate normal prior.

Suppose $y \stackrel{d}{=} \mathcal{N}(X\beta, \Sigma)$, Σ known.
Take a prior $\beta \stackrel{d}{=} \mathcal{N}(b_0, M)$, where b_0 and M are known
Ten the posterior for β is multivariate Normal

$$\beta \mid y, \Sigma, b_0, M \stackrel{d}{=} \mathcal{N}(b, A) \tag{1.20}$$

$$\begin{aligned} A &= (M^{-1} + X'\Sigma^{-1}X)^{-1} \\ b &= (M^{-1} + X'\Sigma^{-1}X)^{-1}(M^{-1}b_0 + X'\Sigma^{-1}y) \end{aligned} \tag{1.21}$$

which is a matrix-weighted average between b and $(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$, the latter being the classical GLS estimator. If $M^{-1} = 0$, we get the dual results from standard generalized regression.

One of the things we can do with (1.21) is explore the effect of prior ideas in a more flexible way than classical testing which uses fixed prior ideas about

β. In econometrics, where regression is often used in a context where many explanatory variables with strong multicollinearity are possible, the strategy of throwing away everything that is not "significant" is more or less standard, and necessary to get reasonable outcomes. By specifying priors centered around plausible values (the values one tests in the classical setup), the resulting outcomes are between the restricted and the unrestricted outcomes.

Note that it is not necessary to run new regressions to compute the results. If the regression output contains the GLS estimate and its covariance matrix, one may put these in a package that allows one to experiment with b_0 and M, to investigate the effect of priors. In this way one can explore what part of the outcomes is robust (not too much dependent on priors, varied around what seems reasonable).

A clear example of the difference between testing and using priors is in Tol and de Vos (1993). They make an analysis of the enhanced greenhouse effect with a model for the global mean temperature (GMT) in the last century. Explanatory variables are the atmosferic concentration of CO_2 (as a proxy for all greenhouse gases and in the form of a weighted average of past values to account for lags), various climatological variables like volcanic activity etc. and a linear trend. The regression outcome

$$GMT = \alpha + \beta_1 \ln(CO_2) + \beta_2 TREND + \textit{ other explanations}$$

has a resulting regression estimate of β_2 which is not significant (a t-value of about 0.5), so a standard procedure would be to delete $TREND$. However, $TREND$ stands for the natural variability in GMT that we cannot explain. Estimates of the movements of GMT in the past 10.000 years show that natural variability occurs: a rise of 0.5 degrees Celsius in one century without an increase in greenhouse gasses is rare, but does occur. From the past records one can make a prior for β_2 : the probability distribution of a prediction from data up to 1890 about the "natural" rise of GMT. There was no reason to expect such a rise, so the prior mean is about zero. The prior standard error would be somewhere between 0.12 and 0.24.

The resulting estimates for β_1 scaled to represent the effect of doubling the present value of greenhouse gasses (to be expected in about a century) are in table 6.1.

Table 6.1 *Effect of the prior for the natural variability parameter β_2 on the posterior for the "Greenhouse effect" parameter β_1*

prior $S.E.[\beta_2]$	∞	0.24	0.12	$0(\beta_2 \equiv 0)$
$E[\beta_1 \mid data]$	4.40	4.17	3.83	3.37
$S.E[\beta_1 \mid data]$	1.45	1.30	1.05	0.28

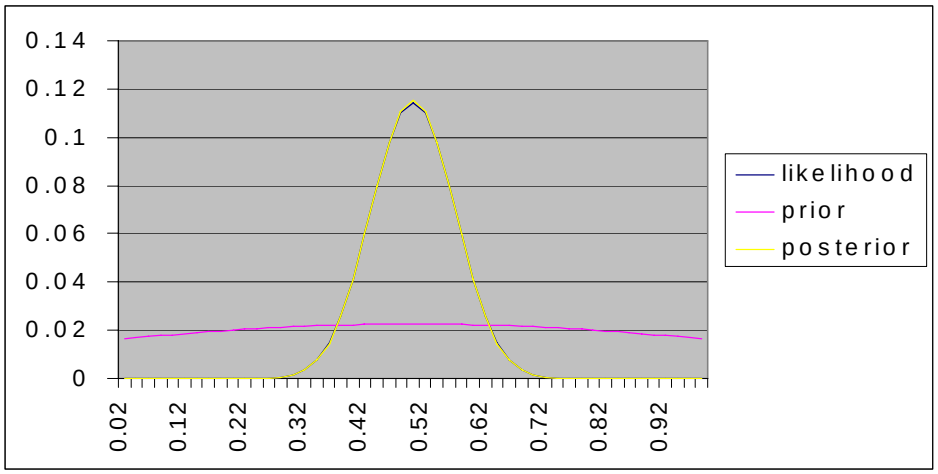
Deleting $TREND$ (the last column) would attribute all rise in GMT to CO_2, resulting in a very low standard error of $\beta_1(0.28)$. Doing just regression is the same as making the prior standard error of β_2 infinite ($S.E.[\beta_2] = \infty$, the first column), but this is rather unrealistic. The outcomes in the middle are the reasonable ones. That the prior for β_2 is so important for the inference on β_1 is due to multicollinearity: CO_2 and $TREND$ are both increasing series, and competing explanations for the rise in GMT.

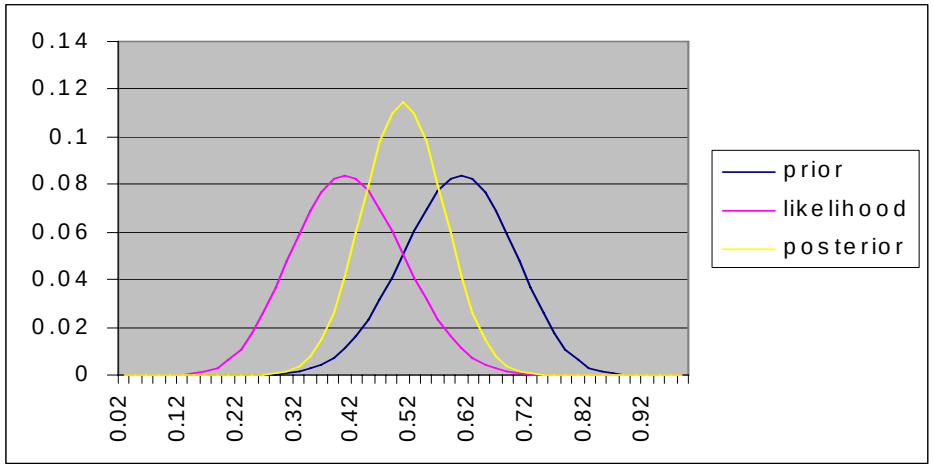
1.6 Visual Bayes

There is no general analytic way to combine information from priors and likelihoods. One has to do it.

If the analysis concerns one parameter, it is easy to interpret from pictures what is going on. Pictures are easily made in a spreadsheet.The most informative picture shows the prior and the scaled likelihood. Scaled such that it integrates to one. It looks a probability distribution then, but don't forget that in general it cannot be interpreted as such.

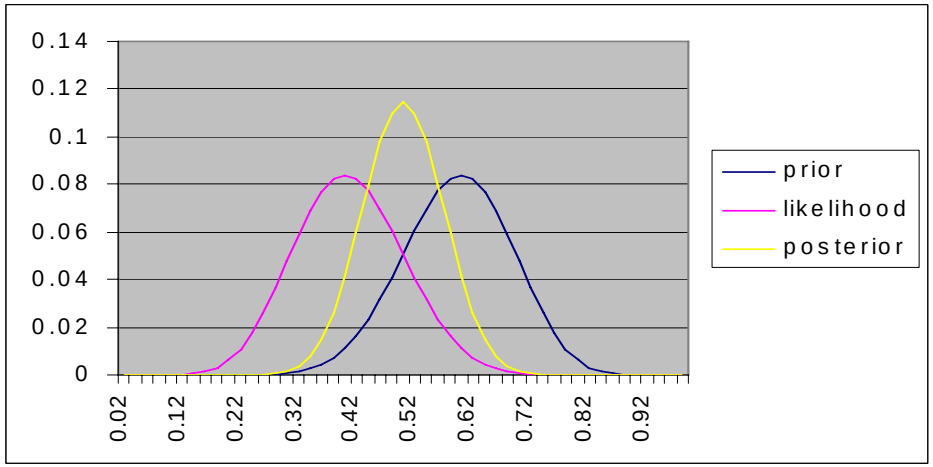
In many cases the likelihood dominates the prior:



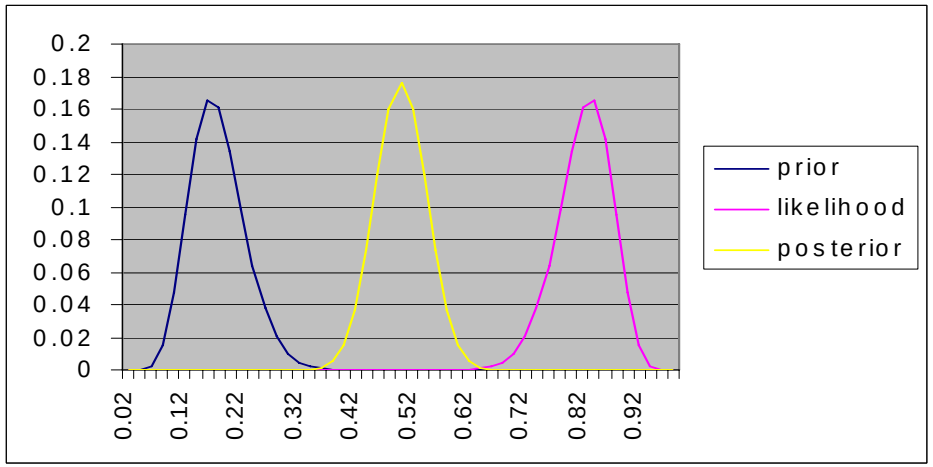


The posterior is almost equal to the scaled likelihood. This is the case where classical and Bayesian inference often reach similar conclusions.

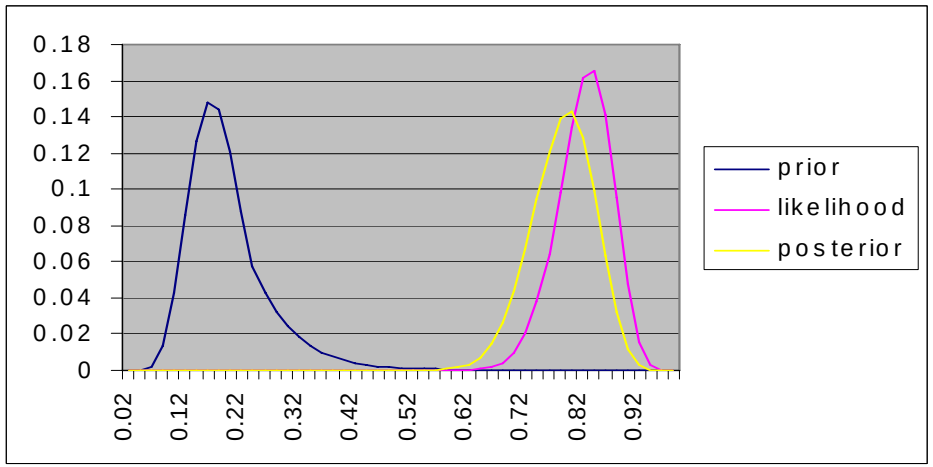
Information from prior and likelihood are both important in pictures like:



Also two sources of information may be in conflict. This situation requires great care. There is something wrong with either the prior or the likelihood and if not, the result is very sensitive for the form of the prior. Normal priors and likelihoods give



But notice the difference between the case above and the result if a fat "maybe I am wrong" tail is attached to the prior:



This prior makes that the likelihood dominates in the case of unexpected outcomes, as is easily seen.

1.7 Elegant pooling of information: conjugate priors

Bayes' formula for combining information on a parameter from a prior and a likelihood will often lead to cumbersome formulae. The choice of an appropriate functional form for the prior may circumvent this problem. Typically, a prior with the same functional form as the likelihood (which is a function of the parameters!) is often suitable. If the resulting posterior distribution has this functional form as well, such a prior is called a conjugate prior. An example:

Poisson likelihood and Gamma prior.

The Poisson distribution is

$$P(x \mid \mu) = \frac{e^{-\mu}\mu^x}{x!} \tag{1.22}$$

If we have data from a Poisson distribution, we use this equation as the likelihood: for fixed x and different μ. As a function of μ, (1.22) looks like a Gamma distribution. Note that one finds this distribution in most books (and in appendix A) as a function of x: if x is distributed $Gamma(r, \phi)$ then

$$f(x) = \Gamma(r)^{-1}\phi^r x^{r-1} e^{-\phi x} \tag{1.23}$$

with expectation r/φ and variance r/φ^2. The functional form of (1.23) as a function of x corresponds with (1.22) as a function of μ. As a conjugate prior for μ we may thus take

$$\pi(\mu) = \Gamma(r)^{-1}\phi^r \mu^{r-1} e^{-\phi\mu} \tag{1.24}$$

where r and ϕ must be chosen such that (1.24) reflects the prior beliefs on μ. As prior beliefs seldom may be specified more accurately than an expectation and a variance, (1.24) is sufficiently rich for most cases.

The convenience of conjugate priors appears when we combine the prior (1.24) and the likelihood (1.22). The product of both is, *deleting all multiplication factors not involving* μ (because they may be included in the proportionality constant)

$$P(\mu \mid x) = c\mu^{x+r-1} e^{-(\phi+1)\mu} \tag{1.25}$$

c simply follows from the fact that the function must integrate to 1. $\mu^{x+r-1}e^{-(\phi+1)\mu}$ is called the **kernel** of the distribution. In this case it is not even necessary to calculate c: by comparing with (1.23), it is clear that $\mu \mid x$ has a $\Gamma(x + r, \phi + 1)$ distribution. This is the nice feature of conjugate priors: the posterior comes from the same family as the prior. Moments can simply be looked up; in this case it is directly clear that

- $E[\mu \mid x] = \frac{x+r}{1+\phi}$, a weighted average of x and the prior expectation $\frac{r}{\phi}$.
- $Var[\mu \mid x] = \frac{x+r}{(1+\phi)^2}$ which is smaller than the variance of the prior $\left(\frac{r}{\phi^2}\right)$, as is easily checked.

In the case of the normal prior and the normal likelihood (another case of conjugacy), we also saw that a variance of the posterior which is always lower than the variance of the prior. In general we expect this; the well known formula

$$Var(x) = E_y\left[Var(x \mid y)\right] + Var_y\left[E(x \mid y)\right] \tag{1.26}$$

applies in the form

$$E_x[Var(\theta \mid x)] = Var(\theta) - Var_x(E(\theta \mid x) < Var(\theta) \tag{1.27}$$

This says that we may expect the posterior variance to be smaller but not that this is certain. If the information from the prior and the likelihood conflict a higher variance of the posterior may result.

An example with the posterior variance higher than the prior variance:

The model $P[X = 1] = \theta$ with prior $\pi(\theta = 0.1) = 0.9$ and $\pi(\theta = 0.9) = 0.1$ gives, combined with an observation $X = 1$, a posterior

$P(\theta = 0.1|X) = P(\theta = 0.9|X) = 0.5$,

which has larger variance. This is due to the "unexpected" outcome.

One would intuitively expect that a conflict between prior ideas and sample evidence would increase uncertainty in general. This is not true in most cases where conjugate priors are used. This is a reason to *be careful with conjugate priors*. The least one can do is to make a picture of prior, likelihood and posterior. If likelihood and posterior differ strongly, one must inspect the prior with more care. It might well be that good priors have fatter tails, to reflect the idea that one may be wrong. Fat tails imply that if the likelihood is concentrated in a tail area, the prior looses its importance.

Exercise 7. A binomial likelihood has a beta prior as a conjugate Derive the formula for the posterior expectation and variance; check whether the variance always decreases.

Think of relevant priors for the following situations:

- breaking a leg on a ski holiday

- surviving a risky operation

-throwing a coin.

And compute posterior probabilities in case you have a sample with 3 "successes" in 10 trials.

1.8 Bayesian Inference with non-informative priors.

In section 5B we noted a surprising duality between classical and Bayesian confidence intervals in the case of the normal distribution. If the prior for μ is uniform on a large enough range the "confidence intervals" coincide, except in interpretation. This "coincidence" has inspired many statisticians to mimic classical statistics in a Bayesian way. A uniform prior for μ on whatever interval is relevant comes close to the idea of "knowing nothing" about μ. This kind of prior is therefore called "noninformative". It is intuitively plausible that the resulting inference resembles the answer to the classical question "what kind of statement about μ is valid whatever the value of μ" but that the answers coincide is -as we will show- due to special circumstances.
Noninformative priors have a special appeal. One might say that statisticians should confine themselves to revealing what the data tell and leave others to combine this with prior information. If there is consensus about the model to be used, this argument is a strong one. As soon as the statistician chooses the model out of many possible models, prior information can no longer be avoided: the set of models to be considered and how choice between them is made involves judgement. In this situation, the use of informative priors is simply a way of incorporating judgement.

Despite the doubts of "real Bayesians", noninformative priors should be considered as an interesting alternative to the classical approach. At any rate, it provides a framework which cannot be accused of subjectivity. It is nowadays a well-established branch of statistics. This is greatly due to the seminal work of Box & Tiao(1973)-henceforth B&T. This book gives an excellent survey of the possibilities for solving classical statistical inference problems with Bayesian methods without or nearly without using prior information, following Jeffreys'(1939, 1961) device to construct prior-distributions.

A. Improper priors

Let us consider the noninformative prior more closely. In the case of the mean μ of the normal distribution, with σ known, we saw in section 5 that the limiting posterior distribution as the prior variance (σ_0^2) goes to infinity resulted in

$$\mu \mid x, \sigma \stackrel{d}{=} \mathcal{N}(\bar{x}, \frac{\sigma^2}{n}) \tag{1.28}$$

We would have obtained the same answer using the prior

$$\pi(\mu) \propto c \tag{1.29}$$

This is not really a probability distribution, it is called an **improper prior. However,** it works, and directly results in

$$P(\mu \mid x) \propto P(x \mid \mu) \tag{1.30}$$

which would allow us to use the likelihood function directly to construct posteriors.

B. A problem with transformations.

It would be tempting to always use priors like (1.29). But there is a little snag which becomes clear by studying the distribution of $\sigma \mid x$: should we say

$$P(\mu, \sigma) \propto c$$

leading to

$$P(\sigma \mid x, \mu) \propto \sigma^{-n} \exp\left\{-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\right\} \tag{1.31}$$

or should we say

$$P(\mu, \sigma^2) \propto c \tag{1.32}$$

This makes a difference: if (1.32) were right then instead of (1.31) we would get:

$$P(\sigma \mid x, \mu) \propto \sigma^{-(n-1)} \exp\left\{-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\right\} \tag{1.33}$$

the extra σ coming from the transformation [1] from σ to σ^2. In the classical setup it does not matter what parameterization we choose: σ, σ^2, $\ln(\sigma)$ or any other monotonic transformation.

There are arguments, going back to Jeffreys, to take

$$P(\mu, \sigma) \propto \sigma^{-1} \tag{1.34}$$

which implies the beautiful result below.

C. The normal distribution with μ and σ unknown.

An interesting application of **??** is the important case of the normal distribution with μ and σ unknown. The noninformative prior is σ^{-1}; multiplication with the likelihood gives the simultaneous posterior for both parameters:

$$P(\mu, \sigma \mid x) \;\propto\; \sigma^{-(n+1)} \exp\left\{-0.5\frac{\sum_i (x_i - \mu)^2}{\sigma^2}\right\} \tag{1.35}$$

So one can draw simultaneous (plausible, egg-like) simultaneous confidence intervals, a task which is very difficult in the classical setup.
If one is only interested in μ, (σ is a "nuisance parameter") the solution is simply to integrate σ out. Using the result from calculus that

$$\int x^{-(p+1)} e^{-ax^{-2}} dx = 0.5 a^{-p/2} \Gamma(p/2) \tag{1.36}$$

it is easy to derive that

$$P(\mu \mid x) = \int P(\mu, \sigma \mid x) d\sigma \;\propto\; \left(1 + \frac{n(\mu - \overline{x})^2}{(n-1)s^2}\right)^{-n/2} \tag{1.37}$$

[1] Remember: if $y = g(x)$ is monotonous transformation of x then $f_Y = \mid \frac{\partial g^{-1}(y)}{\partial y} \mid f_X(g^{-1}(y))$.

where $\overline{x} = \sum_i \frac{x_i}{n}$; $s^2 = \sum_i \frac{(x_i - \overline{x})^2}{n-1}$.

If

$$f(t) \;\propto\; (1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}} \tag{1.38}$$

t has a student distribution with ν degrees of freedom. Consequently, (1.37) implies that

$$\frac{\mu - \overline{x}}{s/\sqrt{n}} \stackrel{d}{=} Student(n-1) \tag{1.39}$$

which is a perfect analogy to the classical result, though with a different interpretation and a completely different derivation.

Another nice result is the marginal distribution for σ^2: it appears that $\frac{(n-1)s^2}{\sigma^2}$ has a $\mathcal{X}^2_{n-1}$ distribution, leading not only to confidence intervals for σ^2, dual to the classical ones, but also, by direct transformation, to a posterior for σ, which is more important for practical purposes.

1.9 Point Estimation.

Classical inference rests on three pillars: point estimation, confidence intervals and hypothesis testing. In Bayesian inference, the division is not that sharp. Point estimation and confidence intervals give information about a parameter; Bayesians look at the posterior distribution of the parameter, which may be used for point estimation, the construction of confidence intervals or other purposes (like the predictive distribution, see the next section). Hypothesis testing may, as point estimation, be seen as some decision problem to be solved with the posterior, but has many more faces, as the last section will show.

To connect posterior distributions with point estimation requires a loss function. The decision is to report one estimate (t) for a parameter (θ). The action is to minimize the expected loss. The choice of the loss function is in principle a separate problem. Only to get nice analytical results, some loss functions are more suited than others. The best known loss function is

A. Squared Loss, $L_S = (t - \theta)^2$

The decision function is:

$$\min_t E[L_S(t|x)] = \int (t - \theta)^2 P(\theta|x) d\theta$$

Differentiating with respect to t and putting the derivative equal to zero learns:

$$\begin{aligned} t\int P(\theta|x)d\theta &= \int \theta P(\theta|x)d\theta \\ \Rightarrow \quad & t = E[\theta|x] \end{aligned}$$

so t is the **posterior expectation of** θ.

B. Absolute Loss, $L_A = |t-\theta|$

Writing out the two possibilities for the expected loss and differentiating learns:

$$\frac{dE[L_A(t|x)]}{dt} = \int_{-\infty}^{t} P(\theta|x)d\theta - \int_{t}^{\infty} P(\theta|x)d\theta$$

the minimum is obtained for $t =$ **the median of** $\mathbf{p}(\theta|x)$**.**

C. Relative Squared Loss, $L_{RS} = \left(\frac{t-\theta}{\theta}\right)^2$ for $\theta > 0$

This gives $t = \frac{E(1/\theta|x)}{E(1/\theta^2|x)}$. This loss function is important in cases in which one has a decent posterior in terms of θ^{-1} , but not in terms of θ. For instance, if one has a normal posterior for θ , the posterior for θ^{-1} has no moments, so the squared loss and the absolute loss have infinite expectation.

D.Linex Loss $L_L = b[(e^{a(t-\theta)} - (1-a(t-\theta))]$, with $b > 0$ and $a \neq 0$

This is an asymmetric loss-function which gives $t = -\frac{1}{a}\ln(E(e^{-a\theta})$. An elegant result arises for normal posteriors, where a normal posterior for θ with expectation $\overline{x}$ and variance $\frac{\sigma^2}{n}$ implies

$$t = \bar{x} - a\frac{\sigma^2}{2n}.$$

Exercize 8: prove this.

Perhaps the most important point in the Bayesian approach is that **the loss function may be applied separately from the computation of the posterior**. Different purposes involve different loss functions, so if it is not clear for what decision your estimate is going to be used, it is better to give the posterior completely. This is especially important for decisions with asymmetric loss functions, which appear frequently in practice; optimal point estimates are easily derived from the posterior, and they cannot be derived from other point estimates.

Bayes vs. Berkeley in point estimation.

The posterior expectation may be considered as the dual solution to the classical Minimal Mean Squared Error (MMSE) estimate. In classical inference, no solution exists for uniformly (whatever the parameters) optimal MMSE estimation. What does exist in some cases are Uniformly Minimum Variance Unbiased Estimates (UMVUES). The condition of unbiasedness ensures solutions in cases where as many sufficient statistics as parameters are involved, and the distribution function belongs to the exponential family. Where UMVUES exist, Bayesian posterior expectations are either coinciding, or slightly deviating in a plausible way. However, the Bayesian solution is possible in any case, has a nicer interpretation (it does not refer to other samples that could have arised but didn't), and is easily adapted to other loss functions.

In classical "unbiased" inference, nonlinear transformations of the parameters create some problems: one cannot simply transform the estimates; new estimators must be derived to obtain unbiasedness. Estimating σ (instead of σ^2) in a Normal distribution is a well known example. In Bayesian inference the posterior of a transformed parameter brings no conceptual problems along, it is simply a matter of transforming the posterior (which does not mean, by the way, that the resulting estimate has an easy analytical expression).

A practical tip at the end: there is no reason why you should report your estimates in the same form as they appear in your distribution, which is often nothing more than a mathematical convention. (examples: in the exponential distribution, $f(x) = \lambda e^{-\lambda x}$, λ^{-1} is the expectation and of more interest than λ; in the normal distribution inference is made on σ^2, while σ is more important).

All together the list of advantages of the Bayesian approach over classical point estimation is impressive:

- no separation of asymptotic and small sample properties,
- no problems with nonlinear transformations,
- separation of statistical inference and loss functions,
- no dependency on sufficient statistics,

- coherency with other aspects of inference, like confidence intervals (in Classical Statistics point estimators, confidence intervals and hypothesis testing are separate items),
- no dependency on the existence of moments.

To illustrate the last point: the Bayesan approach provides ways to do proper inference in difficult cases, like models involving infinite variances, or even nonexisting expectations. An example is the Cauchy distribution

$$f(x) = (\pi\beta)^{-1}(1 + (\frac{x-\alpha}{\beta})^2)^{-1}$$

Some standard notions from statistical inference no longer apply. For instance, the distributions of the mean of a sample is the same as that for a single observation. Still Bayesian inference is possible, along the same line as all other inference.

Exercize 10: Let the sample -10,2,8,45 be a sample from a Cauchy distribution, with $\beta = 1$. Compute and draw in a spreadsheet the prosterior for α with the noninformative prior and compare this with the posterior for the "double sample", where all the values occur twice. What loss functions do you think lead to solutions for the point-estimation of α?

1.10 The predictive distribution.

If the goal of science is to make a prediction of the future given the past, Bayesian statistics is the way to do it. Unlike classical statistics, the Bayesian approach offers the possibility to derive the distribution of a forecast, z. That is, the unconditional distribution of a forecast, i.e. not conditional upon a parameter. To avoid confusion with the standard that the distribution of a forecast is the conditional distribution, the term "predictive distribution" is used.

To come to a predictive density, a model is required.

If z is a future datapoint of a sample, where x are the known data, $P(z|\theta)$ is known by assumption. The result of Bayesian inference is $P(\theta|x)$. This is what one knows about θ, given the sample-results. As

$$P(z|x) = \int_\theta P(z|\theta, x)P(\theta|x)d\theta \tag{1.40}$$

which is simply the formula to get a marginal distribution out of a simultaneous distribution, with x everywhere as an extra conditioning factor. Now

$$P(z|\theta, x) = P(z|\theta) \tag{1.41}$$

because, *once we know θ, x gives no new information about z* (This type of argument is very important in the Bayesian analysis of complex information structures).

So we have the result

$$P(z|x) = \int P(z|\theta)P(\theta|x)d\theta \tag{1.42}$$

The predictive distribution is a weighted average of forecast distributions conditional upon θ, with the posterior for θ as weights. A similar result may only in special cases be obtained in a classical context. The predictive distributions are often new distributions with plausible properties.

Example 1. A binomial distribution with parameter θ gives a funny result:

$$P(z = 1) = \int P(z = 1|\theta)P(\theta|x)d\theta = \int \theta P(\theta|x)d\theta = E[\theta|x]$$

which simply the expectation of the posterior. Note that this means that "being uncertain about the probability θ", in the sense that one has a –prior or posterior– distribution for θ, cannot be distinguished from being certain; each distribution with the same expectation has the same implication for the prediction.

Example 2. A sample consists of values (1 1 2 3 4 5). Assume a noninformative prior. This time we assume that these data come from an exponential(λ) distribution ($\lambda > 0$):

$$P(x|\lambda) = \lambda e^{-\lambda x}$$

or, equivalent and more convenient:

$$P(x|a) = \frac{1}{\theta}e^{-\frac{x}{\theta}}.$$

The noninformative prior for θ is $\pi(\theta) \propto \frac{1}{\theta}$ (θ is scale parameter). So, in the original λ-representation:

$$\pi(\lambda) \propto |\frac{\partial\theta}{\partial\lambda}|\lambda \propto \lambda^{-2}.\lambda \propto \lambda^{-1}$$

Further we have:
Likelihood:

$$P(x|a) = \lambda^n e^{-\lambda \sum_i x_i} = \lambda^6 e^{-16\lambda}$$

Posterior:

$$P(\lambda|x) = c\lambda^{n-1} e^{-\lambda \sum x}$$

with $c^{-1} = P(x) = \int_0^\infty \lambda^{n-1} e^{-\lambda \sum x} d\lambda = (\sum_i x_i)^{-n}\Gamma(n)$. (This is the Gamma integral)

So, as $\sum x = 16, n = 6$,

$$P(\lambda|x) = 16^6 \lambda^5 e^{-16\lambda} \frac{1}{5!}$$

Predictive Density (check this):

$$P(z|x) = \int P(z|\lambda)P(\lambda|x)da = n\left(\sum_i x_i\right)^n \left(\sum_i x_i + z\right)^{-(n+1)}.$$

Again we obtain plausible new distributions. A good check is to control whether the limiting PD is of the same form as the likelihood:

$$\lim_{n\to\infty} \frac{n(\sum x)^n}{(\sum x + z)^{n+1}} = \lim_{n\to\infty} n(1 + \frac{zn}{n\sum x})^{-n}/(n\frac{\sum x}{n} + z) = \lambda e^{-\lambda z},$$

as

$$\lim_{n\to\infty} \frac{\sum x}{n} = \frac{1}{\lambda},$$

and

$$\lim_{n\to\infty} \frac{n}{z + n/\lambda} = \lambda,$$

$$\lim_{n\to\infty} (1 + \frac{z\lambda}{n})^{-z\lambda(n/z\lambda)} = e^{-z\lambda}.$$

Exercize 11: simulate for this example a predictive distribution: draw the parameter from the posterior and next the future value with this parameter (2 cells in a spreadsheet!). Compare with the analytical result.

Exercize 12: Derive the predictive distribution in a sample from $\mathcal{N}(\mu, \sigma^2)$,

A. for μ unknown, σ known (a derivation similar to the computation of the posterior).

B μ and σ unknown; use noninformative priors.

Chek that
$Var(z|x) = E_\theta Var(z|x,\theta) + Var_\theta E(z|x,\theta)$

1.11 Bayesian Inference on Hypotheses.

Hypothesis testing is a much more controversial issue than inference on parameters. Classical and Bayesian solutions differ strongly. The choice of parameter restrictions and the choice between different models is, especially in econometrics, the domain where hypothesis tests are used. The classical "only use large models when the outcomes are unlikely under smaller models" has a Bayesian counterpart where posterior probabilities of models are computed. Many articles on "Bayes Factors" have recently appeared. The Bayes Factor transforms prior beliefs in models into posterior beliefs. In the case of vague priors it favors small models. Some see this as a natural advantage of Bayes factors, others do not like the nonrobustness with respect to prior vagueness. An extreme case is the noninformative prior. This case, where in parameter estimation Bayes and the classical setup come close, leads to trouble in the Bayes Factor: larger models are always rejected. Recent research has led to interesting new discussions, but they are beyond the reach of this book. We confine ourselves here to the "traditional Bayesian" theory.

A. Hypotheses as extensions of the model space.

A hypothesis is an assumption about a possible state of nature. If models are hypothesized with known parameters, each hypothesis is simply a complete statement about a possible probability distribution for the data. One may specify hypotheses that belong to the same model and only differ with respect to the parameters, e.g., $\mathcal{N}(-1,1)$, $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$ as three possibilities, but one may just as well take hypotheses that come from different models, say $\mathcal{N}(-1,1)$, Sech(0,1) and Cauchy(1,1). In the Bayesian setup these possibilities are equivalent, and inference is possible once one assigns prior possibilities to all hypotheses. So, there is no strict separation between "different models" and "different parameters for the same model". All hypotheses are simply competing descriptions; prior probabilities must be attached, and posterior probabilities will result. If one has hypotheses $H_1 ..... H_n$, Bayes' rule says:

$$\begin{aligned} P(H_i \mid x) &= \frac{P(x \mid H_i)\pi(H_i)}{P(x)} \\ &= \frac{P(x \mid H_i)\pi(H_i)}{P(x \mid H_1)\pi(H_1) + \ldots + P(x \mid H_n)\pi(H_n)} \end{aligned} \tag{1.43}$$

with $\pi(H_1) + \ldots + \pi(H_n) = 1$. An alternative notation is:

$$P(H_i \mid x) \propto P(x \mid H_i)\pi(H_i) \tag{1.44}$$

which stresses the fact that the two inputs are again prior probabilities and likelihoods. If only two models, H_0 and H_1 are involved, another equivalent, even more elegant, representation is in terms of "odd ratios":

$$\frac{P(H_0 \mid y)}{P(H_1 \mid y)} = \frac{\pi(H_0)}{\pi(H_1)} \cdot \frac{P(y \mid H_0)}{P(y \mid H_1)} \tag{1.45}$$

or: *the posterior odds are the prior odds times the likelihood-ratio.* "odds $\frac{a}{b}$" is, in the case of two alternatives the same as probabilities $\frac{a}{a+b}$ and $\frac{b}{a+b}$. One may also express (1.43) in terms of odds, but then the link to probabilities is less obvious.

The essential point is that the hypotheses are treated stochastic, just like parameters. For fully specified hypotheses the formulae may directly be used. It is also possible to specify different parameterized models H_i , each model having parameter(s) θ_i. If prior probabilities $\pi(H_i)$ are given, together with priors for the parameters of any model $\pi(\theta \mid H_i)$, one comes in the preceding situation by using the **"marginal likelihood"**:

$$P(x \mid H_i) = \int_\theta P(x \mid H_i, \theta_i)\pi(\theta_i \mid H_i)d\theta_i \tag{1.46}$$

for all models.

The results are posterior probabilities for the hypotheses, $P(H_i \mid x)$, together with posteriors for the parameters in all of the models, $P(\theta_i \mid H_i, x)$. The ratio of the marginal likelihoods for two models is called the **Bayes Factor**.

B. Forecasting with mixed models.

In forecasting one may simply proceed with the mixture of models one has obtained. Let z be the following observation. If the models are completely specified, so parameters are only involved in known form in the different models then

$$P(z \mid x) = \sum_i P(z \mid H_i)P(H_i \mid x) \tag{1.47}$$

just like (9.3), hypotheses replacing parameters.

If parameters are involved, a beautiful schedule involving models, parameters, data and forecasts results, simply by application of the laws of probability calculus. First (9.3) may be used conditional upon each model:

$$P(z \mid x, H_i) = \int P(z \mid \theta_i, H_i)P(\theta_i \mid x, H_i)d\theta_i \tag{1.48}$$

$P(x \mid H_i)$ is obtained directly as a by-product in the calculation of $P(\theta_i \mid x, H_i)$, it is the scaling factor of $P(x \mid \theta_i, H_i)\pi(\theta_i \mid H_i)$. With the priors $\pi(H_i)$, (1.45) provides $P(H_i \mid x)$. Next

$$P(z \mid x) = \sum_i P(z \mid x, H_i)P(H_i \mid x) \tag{1.49}$$

is used (this is simply obtaining the marginal distribution of z, from the simultaneous distribution of z and H, all distributions being conditional upon x). As $P(z \mid x, H_i)$ is known from (1.48), the goal is attained. Note the difference between (1.49) and (1.47): as parameters are involved, $P(z|H_i, x)$ is no longer independent from x, as (1.48) shows.

Exercize 13. Choose 5 models, with known parameters.

A. Generete data from one of the models, and show how the posterior probabilities for the models converge to the "true" model (take prior weights 0,2)

B. Generate data from a mixture of two of the models and show how the posterior weights converge to this mixture.

C. Think about the relevance of the following "ceramelick" (written after a lecture on ceramics which are strong but brittle)

Just study ceramics and it′ll

Tell you why Bayes helps us so little

If the true model lacks

The coherency cracks

So our Theorem is strong, but itıs brittle.

C. Recent developments in Bayes Factors

The huge recent ltterature on Bayes Factors has not led to unanimity among Bayesians on how to perform model choice. In general, vague priors lead automatically to Bayes Factors favoring small models. Unfortunately the Bayes Factors are very sensitive for the vagueness of the prior. Some Bayesians (E.E. Leamer specifically) argue that one should not choose models on their "probability of being true", but on their utility, simplicity being an advantage as such. All Bayesians agree that classical testing procedures are badly motivated, but their alternatives are not (yet) very convincing either.

1.12 The MCMC revolution

From 1990 a revolution took place in Bayesian computing. The main problem in the application of Bayesian methods was until then that computation was difficult. With conjugate priors analytical results could sometimes be obtained, but in complex models even that did not bring much relief. With the increasing speed of computers however, simulation methods became more and more feasible. Methods like "importance sampling ". But in 1990 the "Markov chain Monte Carlo " (MCMC) methods were found. A Markov chain means a simulation where the next draw only depends on the previous one. The technique has in itself nothing to do with Bayesian analysis. The question is whether it is possible to simulate from a complex distribution. As in Bayesian inference the posterior is known if prior and likelihood are known it is possible to write down the formula for it. And if there is a trick to simulate given the formula, that is sufficient.

A number of MCMC methods performs this trick. The basis of the simulation are proofs that a simulation scheme provides -possibly after some time- drawings from the programmed distribution.

There are two basic schemes for which such a proof has been given. The "Gibbs Sampler " and the "Metropolis-Hastings algorithm ".

The Gibbs sampler goes as follows:

- Suppose there are parameters $\theta_1.....\theta_n$ and that the distribution of each θ_i conditional upon the other $\theta's$ is known.
- Draw in each step $\theta_i|\theta_1..\theta_{i-1},\theta_{i+1}..\theta_n$, and cycle through θ_i (so $\theta_1,\theta_2...\theta_n,\theta_1,\theta_2.....$

This is a very efficient MCMC algorithm but the requirement that the conditional distribution must be known requires in many cases a lot of work and in even more cases it is not analytically possible, the main problem being the normalizing constant.

With **the Metropolis-Hastings** algorithm life becomes really simple. Only the formulas for the likelihood and the prior must be programmed, it is not even needed to know the normalizing constant of the posterior.

The easiest setup for the one dimensional case is given in the accompanying spreadsheet "metropolis.xls". The goal is to simulate from $f(\theta)$, in this case the function $\exp(-|\theta|)$, so, apart from the normalizing constant, a double exponential (or Laplace) distribution for θ. The algorithm consists of the simple steps (θ_i denoting draw i):

- 1. Start with some θ_1 ($i = 1$)
  2. The proposal for $\theta_{i+1}, \theta^p_{i+1}$ is a draw from uniform$(\theta_i - a, \theta_i + a)$
  3. If $f(\theta^p_{i+1}) > f(\theta_i)$ then the proposal is accepted: $\theta_{i+1} := \theta^p_{i+1}$; else the proposal is accepted with probability $\alpha = f(\theta^p_{i+1})/f(\theta_i)$. If not accepted, $\theta_{i+1} = \theta_i$

Go back to step 2.

One can easily check for a discrete distribution that if θ_i is a draw from $f(\theta)$, then so is θ_{i+1}. This holds also for continuous distributions.

There are many variations of the MH algorithm. This was the "random walk " variant with uniform proposals. One may also do random walk proposals with draws from other distributions that ar symmetrical around θ_i. If this distribution is $g(\theta)$, the necessary adjustment is to use $\alpha = f(\theta^p_{i+1})/f(\theta_i) \times g(\theta)/g(\theta^p_{i+1})$, to adjust for the probability of the draw.

One may also draw from proposal distributions that are not centered around θ_i. Each time form the same distributions, or adapting the drawing mechanism in some way to make it more efficient. The main requirement for the mechanisms to provide drawings from the posterior is the requirement that the probability to jump from θ_i to θ_{i+1} must be equal to the probability to jump from θ_{i+1} to θ_i. The "reversable jump " property. What is the most efficient way in a highly multivariate context is a science on its own. The random walk variants seem to be winning.

1.13 WINBUGS

Winbugs is the ultimate MCMC program, freely available from the internet. You only have to specify your likelihood and your priors (not the formulas but just the names, unless you want to do very special things), winbugs does the rest.

To see what a student prior with 3 degrees of freedom and a normal likelihood give as posterior (like in the visual Bayes section, one specifies:

```
Model
{
y~dnorm(m,1)
m~dt(0,1,3)
}
data
list(y=3)
```

Updates 20.000 times (10 seconds on my slow laptop) and gets the (almost normal) posterior for m:

node	*mean*	*sd*	*MCerror*	2.5%	*median*	97.5%
m	1.987	0.956	0.0096	0.26	1.943	3.987

Exercize: see what happens if the prior is normal and check what you know analytically

1.13.1 A hierarchical Poisson model

The hierarchical model:

$\ln(\lambda_i) = X\beta + u_i$

$Y_i = Poisson(\lambda_i)$

Can be applied to many cases. Car accidents, Number of medals for a country during the olympics, number of doctor visits. There is no really satisfactory frequentist way to estimate the model. The "Poisson regression model ($Var(u_i) = 0$) can be estimated by maximum likelihood, but is very unrealistic. There is a solution called "Poisson regression with overdispersion " but that is ugly. The hierarchical Poisson model is the natural specification.

In this model not only the parameters β are simulated during the MCMC runs but the latent u_i as well. This possibility to simulate latent variables makes the algorithms incredably versatile.

The winbugs code for this model (here with doctor "visits " explained from 4 explanatory variables),(text after # is comment):

```
model
{
for(i in 1:n) {
visits[i] ~dpois(mu[i])
error[i] ~dnorm(0, tae)
mu[i] <- exp(b[1] + b[2]*x1[i] + b[3]*x2[i] + b[4]*x3[i] + error[i])
}
for(j in 1:4) {
b[j] ~dnorm(0, 0.01)#vague prior for regression coefficients
}
tae ~dgamma(0.01, 0.01)#vague prior for precision tae
}
```

1.13.2 An Unobserved component trend-Poisson model

The model

$y_t = Poisson(\lambda_t)$

$\ln \lambda_t = a_t + u_t$

$a_t = a_{t-1} + \varepsilon_t$

with u_t and ε_t normally (or student) distributed, may describe nicely e.g. sales figures. Now u_t as well as ε_t are simulated during the algorithm.

The Winbugs code is simple :

```
model
{
y[i] ~dpois(lab[i])
lab[i] ~dlnorm(a[i],tau)
tau ~dgamma(0.01,0.01) #vague prior for precision tau
tae ~dgamma(0.01,0.01)
a[1] <-start
start ~dnorm(0,0.0001)#vague prior for d[1]
for (i in 1:N) {e[i] ~dnorm(0,tae)}
for (i in 2:N) {a[i] <-a[i-1]+e[i]}
for (i in 1:N) {exa[i] <-exp(a[i])}
#and to predict future values:
af[1] ~dnorm(a[N],tae)
u[1] ~dnorm(0,tau)
labf[1] <-exp(af[1]+u[1])
z[1] ~dpois(labf[1]) #prediction 1 period ahead
for (i in N+2:N+3) {
af[i-N] ~dnorm(af[i-N-1],tae)
u[i-N] ~dnorm(0,tau)
labf[i-N] <-exp(af[i-N]+u[i-N])
z[i-N] ~dpois(labf[i-N]) #more periods ahead
}
}
```

That is all and for not too long datasets it runs quickly. For longer datasets more advanced MCMC algorithms are needed (the "simulation smoother " to avoid slow conversion due to the heavy correlation between successive values of the trend a_t.

The result is directly suited for e.g. inventory problems: simulations from the simultaneous distribution of the future values of y_t (including the uncertainy about the parameters).

Exercise run the program for te dataset (given as needed in winbugs):

list(y=c(5,2,3,2,5,2,3,1,0,1,0,0,1,3,1,7,6,8,3,5,4,3,5,3,2,4,2,1,0,0),N=30)

1.13.3 The Deviance Information Criterion (DIC)

In 2002 Spiegelhalter et al published a seminal article on the DIC criterion. It may be used to choose between models. Winbugs has an option to compute

it.

The criterion is $-2*$loglikelihood evaluated in the mean of the posterior (so it must be minimised), adjusted for the use of degrees of freedom. This is estimated by the difference between the value in the mean and the (lower) mean of $-2*$loglikelihood during the simulations.

It can for instance be used in the previous models to see whether student distributions give a better fit than normal distributions. But for that case there is an alternative: simply estimate the number of degrees of freedom. Best is to do it in both way and to see whether the conclusions match.

Books (* recommended) and review articles.

Bauwens, L; M.Lubrano; J.F. Richard (1999) "Bayesian Inference in dynamic econometric models" Oxford U.P.

*Berger, J.O.(1985) Statistical Decision Theory and Bayesian Analysis. Springer-Verlag.

Bernardo, J.M, and A.F.M. Smith. "Bayesian Theory", Wiley.

*Bernardo, J.M "Bayesian Statistics"(2003). Survey article, last version can be downloaded from the home page of Jose Bernardo.

Box,G.E.P. and G.C. Tiao (1973) "Bayesian Inference in Statistical Analysis". Addison-Wesley.

Jeffreys(1939/1967) "Theory of probability". Oxford university Press.

*Geweke, John(2005) "Contemporary Bayesian Econometrics and Statistics". Wiley, New York.

*Koop,G., Poirier, D.J., Tobias, J.L. "Bayesian Econometric Methods" (Econometric Exercises 7) Cambridge University Press.

Leamer, E.E. (1978), Specification Searches, Ad-hoc Inference with Nonexperimental Data, Wiley, New York.

*Lee, P.M. (2003) "Bayesian Statistics, an introduction, 3nd ed". Edward Arnold, London.

*Lancaster, T.(2004) "an introduction to modern Bayesian econometrics". Balckwell.

Lindley, D.V. (1985) "Making Decisions". Wiley.

Lindley, D.V.(1990) "The present position in Bayesian statistics" (with discussion), Statistical Science, 5, pp 44-89

*O'Hagan, A(1994) "Kendalls Advanced Theory of Statistics 2B: Bayesian Inference" London: Edward Arnold.

*Poirier, D,J (1995) "Intermediate Statistics and Econometrics" MIT Press.

*Pratt,J.W., Raiffa,H. and R.Schlaifer(1995) "Introduction to Statistical Decision Theory". MIT Press.

Press, S.J. (1989) "Bayesian Statistics: Principles, Models and Applications", Wiley.

Savage, L.J. (1954) The foundations of Statistics. Wiley.

Tol, R.S.J, De Vos, A.F (1998) "A Bayesian Analysis of the Enhanced Greenhouse Effect" Climate Change 38, pp 87–112.

Wille, F.J. (2003) "Auditing Using Bayesian Decision Analysis" Dissertation, Free University, Amsterdam.

Zellner, A. (1971) "An introduction to Bayesian inference in econometrics.", Wiley.