VII: BAYESIAN METHODS

1. Classical statistics and its limitations.

Broadly speaking, statistics splits into two main streams: (i) classical, or frequentist, and (ii) Bayesian. Much of classical statistics is devoted to the following general areas: Estimation of parameters (I), Hypothesis testing (II). Again, this is not exhaustive: the main remaining area is Non-parametric statistics (VI). Estimation of parameters itself splits, into
(ia). Point estimation [e.g., maximum-likelihood estimates, MLEs],
(ib). Interval estimation [e.g., confidence intervals].
Both these are open to interpretational objections. A point estimate is a single number, which will almost certainly be wrong [i.e., will differ from the value of the parameter it estimates]. How wrong? And how to proceed?

A confidence interval is more informative, because it includes an error estimate. For instance, its mid-point can be regarded as a point estimate, and half its length as an error estimate – leading to conclusions of the form

$$\theta = 3.76 \pm 0.003 \qquad (*)$$

– with *confidence* 95% [or 99 %, or whatever]. What does this mean? It is not a probability statement:
either θ lies between 3.73 and 3.79 [(*) is true, so holds with pr. 100 %]
or it doesn't [(*) is false, so holds with pr. 0 %].
Problem: We don't know which!
Interpretation. If a large number of statisticians independently replicated the analysis leading to (*), then about 95 % of them would succeed in producing confidence intervals covering the unknown parameter θ. But
(a) We wouldn't know *which* 95 %,
(b) This is of doubtful relevance anyway. The large number of independent replications will usually never take place in practice. So confidence statements like (*) lack, in practice, a direct interpretation. [They are 'what happens to probability statements in classical statistics when we put the numbers in'.]

A further problem is that small changes in our data can lead to abrupt discontinuities in our conclusions. In borderline situations, θ 'just within'

the confidence interval and 'just outside' represent diametrically opposite outcomes, but the data may be very close. Small changes in input *should* only lead to small changes in output, rather than abrupt changes.

Hypothesis testing is open to similar objections. It is usual to have a null hypothesis, H_0, representing our present theory (the 'default option'), and an alternative hypothesis, H_1, representing some proposed alternative theory. At the end of the investigation, we have to choose between two alternatives. We may be wrong: we may
reject H_0 when it is true, and choose H_1 [Type I error, probability α, the significance level], or
reject H_1 when it is true, and choose H_0 [Type II error, probability β].
We then have a trade-off between α and β. It is not always clear how to do this sensibly, still less optimally [it is customary to choose $\alpha = 0.05$ or 0.01, and then try to minimise β, but this is merely conventional]. Again, problems present themselves:
(i) We won't know whether our choice between H_0 and H_1 was correct;
(ii) Small changes in the data can lead to abrupt changes between choosing H_0 and choosing H_1.
Note. This sort of problem occurs everywhere in life, and not just in classical statistics. Think of any really important decision that turns out to be "touch and go": Brexit vote, Trump vote; goals in football (goal-line technology is very helpful here, but doesn't eliminate the need for judgement in borderline cases); offside decisions in football (ditto), line calls in tennis, etc.; degree classification (problems at the margin between I/II-1, II-1/II-2 etc.; criminal trials ('innocent till proved guilty'; accused 'gets the benefit of the doubt'; beyond all reasonable doubt – how reasonable is reasonable?, etc.

Thus both the main branches of classical parametric statistics lead to abruptly discontinuous conclusions and present interpretational difficulties. One justification for Bayesian statistics is that it avoids these. There are many others: we shall argue for Bayesian statistics below on its merits.

2. Prior knowledge and how to update it.

The difficulties identified above arise because in classical statistics we rely *entirely* on the data, that is, on the sample we obtained. The mathematics involved in classical statistics amounts to comparing the sample we actually obtained with the large (usually, infinite) class of hypothetical samples we might have obtained but didn't. These include the samples that we would

obtain if we repeated our sampling independently – or that other statisticians would obtain if they independently replicated our work. This is where the term 'frequentist' for classical statistics originates: e.g., in 95 % confidence intervals, independently replicated confidence intervals would cover the parameter θ with frequency 0.95.

The other aspect of classical statistics crucial for our purposes is that it ignores everything before sampling. This is often unreasonable. For instance, we may know a good deal about the situation under study, based on prior experience. Such situations are typical in, e.g., industrial quality control: suppose we are employed by a rope manufacturer, and are testing the breaking strain of ropes in a current batch. We may have to hand large amounts of data obtained from tests on previous batches from the same production line. Similarly for a scientist, testing a scientific hypothesis.

In hypothesis testing, such prior knowledge by the experimenter (scientific, manufacturing etc.) is tacitly assumed, because we need it to be able to formulate H_0 and H_1 sensibly. But we may not be willing to enter the 'accept or reject' framework of hypothesis testing [which some statisticians believe is inappropriate and damaging]: how then can we use prior knowledge? In the estimation framework also, we may know a lot about θ before sampling [as in the rope example above]: indeed, if we do *not* have some prior knowledge of the situation to be studied, we would in practice not have enough prior interest in it to be willing to invest the time, trouble and money to study it statistically.

Bayesian statistics addresses this by giving a framework where
1. The statistician knows something before sampling: he has some *prior knowledge*.
2. He then draws a sample, and analyses the *data* to extract some relevant information.
3. He then *updates his prior information* with his *data (or sample) information*, to obtain *posterior information*

(prior: before (sampling); posterior: after (sampling)).

This verbal description of the Bayesian approach is attractive, because it resembles how we learn – and indeed, how we live. Life involves (indeed, largely consists of) a constant, ongoing process of acquiring new information and using it to update our previous ('prior') information/beliefs/attitudes/policies.

To implement the Bayesian approach, we need some mathematics. The formulae below derive from the work of the English clergyman

Thomas BAYES (1702-1761): *An essay towards solving a problem in the doctrine of chances* (1763, posth.).
Recall that if A, B are events of positive probability,

$$P(A) > 0, \qquad P(B) > 0,$$

the *conditional probability* of A *given* (or *knowing*) B is

$$P(A|B) := P(A \cap B)/P(B).$$

Symmetrically,

$$P(B|A) := P(B \cap A)/P(A) = P(A \cap B)/P(A).$$

Combining,

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

$$P(B|A) = P(A|B)P(B)/P(A) \qquad \text{(Bayes' formula, or Bayes' theorem).}$$

Interpretation.
1. Think of A as a 'cause', B as an 'effect'. We naturally first think of P(effect B|cause A). We can use Bayes' formula to get from this to P(cause A|effect B) (think of B as an effect we can see, A as an effect we can't see).
2. Suppose we are interested in event B. We begin with an initial, *prior* probability $P(B)$ for its occurrence. This represents how probable we initially consider B to be [this depends on us: we will have to estimate $P(B)$!]. Suppose we then observe that event A occurs. This gives us new information, which affects how probable we should now consider B to be, *after* observing A [or, to use the technical term, *a posteriori*]. Bayes' theorem tells us how to do this updating: we multiply by the ratio $P(A|B)/P(A)$:

$$P(B|A) = P(B).P(A|B)/P(A) :$$

posterior probability of B = prior probability of B $\times$ updating ratio.
We first observe some extreme cases.
Independence. If A, B are independent, $P(A \cap B) = P(A).P(B)$, so

$$P(B|A) = P(A \cap B)/P(A) = P(A).P(B)/P(A) = P(B),$$

and similarly $P(A|B) = P(A)$: updating ratio = 1, posterior probability = prior probability – conditioning on something independent has no effect.

Inclusion.
1. $A \subset B$: here, $P(A \cap B) = P(A)$, $P(A|B) = P(A \cap B)/P(B) = P(A)/P(B)$;

updating ratio $P(A|B)/P(A) = 1/P(B)$, posterior probability = 1.
2. $B \subset A$: $P(A \cap B) = P(B), P(A|B) = P(A \cap B)/P(B) = P(B)/P(B) = 1$;
updating ratio $P(A|B)/P(A) = 1/P(A)$, posterior probability = $P(B)/P(A)$.
Partitions. B partitions Ω into two disjoint events B; A is the disjoint union of $A \cap B$ and $A \cap B^c$, so

$$P(A) = P(A \cap B) + P(A \cap B^c) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

Similarly, if $\Omega = \cup_1^n B_i$ with B_i disjoint, $A = \cup_1^n (A \cap B_i)$, disjoint. So by finite additivity,

$$P(A) = \Sigma_{r=1}^n P(A \cap B_r) = \Sigma_{r=1}^n P(A|B_r)P(B_r) \quad \text{(Formula of total probability)},$$

using the definition of conditional probability again.

Such expressions are often used for the denominator in Bayes' formula:

$$P(B_r|A) = P(B_r)P(A|B_r)/P(A) = P(B_r)P(A|B_r)/\Sigma_k P(B_k)P(A|B_k).$$

3. Prior and posterior densities.

Suppose now we are studying a parameter θ. Suppose we have data x [x may be a single number, i.e. a scalar, or a vector $x = (x_1, \cdots, x_n)$ of numbers; we shall simply write x in both cases]. Recall that x is an observed value of a random variable, X say. In the *density case*, this random variable has a (probability) *density* (function), $f(x)$ say, a non-negative function that integrates to 1:

$$f(x) \geq 0, \qquad \int f(x)dx = 1$$

(here and below, integrals with limits unspecified are over everything).
Interpretation. $P(X \in A) = \int_A f(x)dx$ for measurable sets $A \subset \mathbb{R}$. For instance, if $A = (-\infty, x]$,

$$F(x) := P(X \in (-\infty, x]) = P(X \leq x) = \int_{-\infty}^x f(y)dy \quad \forall x \in \mathbb{R};$$

as x varies, $F(x)$ gives the (probability) distribution (function) of X.]
In brief: the density $f(x)$ describes the *uncertainty* in the data x.

The distinctive feature of Bayesian statistics is that it treats *parameters* θ in the same way as *data* x. Our initial (prior) uncertainty about θ should also be described by a density $f(\theta)$:

$$f(\theta) \geq 0, \qquad \int_{-\infty}^{\infty} f(\theta)d\theta = 1, \qquad P(\theta \in A) = \int_A f(\theta)d\theta \qquad \forall A \subset \mathbb{R},$$

where the probability on the left is a *prior probability*. The analogue for densities of Bayes' formula $P(B|A) = P(B)P(A|B)/P(A)$ now becomes

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x). \qquad (*)$$

The density on the left is the *posterior density* of θ *given* the data x; it describes our uncertainty about θ knowing x. Now densities integrate to 1: $\int f(\theta|x)d\theta = 1$, so $\int [f(\theta)f(x|\theta)/f(x)]d\theta = 1$:

$$\int f(\theta)f(x|\theta)d\theta = f(x).$$

Combining,

$$f(\theta|x) = f(\theta)f(x|\theta)/\int f(\theta)f(x|\theta)d\theta.$$

In the *discrete case*, θ and/or x may take discrete values $\theta_1, \theta_2, \cdots, x_1, x_2, \cdots$ only, with probabilities $f(\theta_1), f(\theta_2), \cdots, f(x_1), f(x_2), \cdots$. The above formulae still apply, but with *integrals replaced by sums*:

$$P(X \in A) = \Sigma_{x \in A} f(x), \qquad P(\theta \in B) = \Sigma_{\theta \in B} f(\theta),$$

$$f(x) = \Sigma_\theta f(\theta f(x|\theta), \quad f(\theta|x) = f(\theta)f(x|\theta)/\Sigma_\theta f(\theta)f(x|\theta).$$

In the formula $f(\theta|x) = f(\theta)f(x|\theta)/f(x)$, it is θ, the parameter under study, which is the main focus of interest. Consequently, the denominator $f(x)$ – whose role is simply to ensure that the posterior density $f(\theta|x)$ integrates to 1 (i.e., really is a density) – can be omitted (or understood from context). This replaces the *equation* above by a *proportionality statement*:

$$f(\theta|x) \propto f(\theta)f(x|\theta)$$

(here $\propto$, read as 'is proportional to', relates to the variability in θ, which is where the action is). Now $f(x|\theta)$ can be viewed in two ways:
(i) for fixed θ as a function of x. It is then the density of x when θ is the

true parameter value,
(ii) for fixed/known/given data values x as a function of θ. It is then called the *likelihood* of θ (Fisher), familiar from IS, Ch. I, Ch. II, etc.

The formula above now reads, in words:

$$\textbf{posterior} \propto \textbf{prior} \times \textbf{likelihood.}$$

This is the essence of Bayesian statistics. It shows how Bayes' theorem may be used to *update* the *prior* information on θ before sampling by using the information in the *data* x – which is contained in the *likelihood* factor $f(x|\theta)$ by which one multiplies – to give the *posterior* information on θ after sampling. Thus posterior information combines two sources: prior information and data/sample/likelihood information.

4. Examples.

Example 1. Bernoulli trials with Beta prior ([O'H], Ex. 1.4, p.5).

Here θ represents the probability of a head on tossing a biased coin. On the basis of prior information, θ is assumed to have a prior density proportional to $\theta^{p-1}(1-\theta)^{q-1}$ $(0 \le \theta \le 1)$ for $p, q > 0$:

$$f(\theta) \propto \theta^{p-1}(1-\theta)^{q-1} \qquad (0 \le \theta \le 1).$$

Writing

$$B(p,q) := \int_0^1 \theta^{p-1}(1-\theta)^{q-1} d\theta$$

(the *Beta function)*,

$$f(\theta) = \theta^{p-1}(1-\theta)^{q-1}/B(p,q).$$

[We quote the *Eulerian integral* for the Beta function: for

$$\Gamma(p) := \int_0^\infty e^{-x}x^{p-1}dx \quad (p>0), \quad B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q) \quad (p,q>0).]$$

Note that, as p, q vary, the shape of $f(\theta)$ varies – e.g, the graph is u-shaped if $0 < p, q < 1$, n-shaped if $p, q > 1$. Here p, q are called *hyperparameters* - they are parameters describing the parameter θ.

Suppose now we toss the biased coin n times (independently), observing

x heads. Then x is our data. It has a discrete distribution, the binomial $B(n,\theta)$, described by

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \qquad (x = 0, 1, \cdots, n).$$

We apply Bayes' theorem to update our prior information on θ – our prior values of p, q – by our data x. Now

$$f(x) = \int f(\theta)f(x|\theta)d\theta = \int \frac{\theta^{p-1}(1-\theta)^{q-1}}{B(p,q)}.\binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta$$

$$= \binom{n}{x}.\frac{1}{B(p,q)}.\int_0^1 \theta^{p+x-1}(1-\theta)^{q+n-x-1}d\theta = \binom{n}{x}.\frac{B(p+x,q+n-x)}{B(p,q)}.$$

So Bayes' theorem gives

$$f(\theta|x) = f(\theta)f(x|\theta)/f(x) = \binom{n}{x}.\frac{1}{B(p,q)}.\theta^{p+x-1}(1-\theta)^{q+n-x-1}/\binom{n}{x}.\frac{B(p+x,q+n-x)}{B(p,q)}$$

or

$$f(\theta|x) = \frac{\theta^{p+x-1}(1-\theta)^{q+n-x-1}}{B(p+x,q+n-x)}.$$

The posterior density of θ is thus another Beta density, $B(p+x,q+n-x)$.
Summarising:
prior $B(p,q)$ is updated by data x heads in n tosses to posterior
$B(p+x,q+n-x)$.
Graphs. To graph the three functions of θ – prior, likelihood and posterior – first find their maxima.
Likelihood: $f(x|\theta)$ has a maximum where $\log f(x|\theta)$ has a maximum, i.e. where
$x\log\theta + (n-x)\log(1-\theta)$ has a maximum, i.e. where

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0: \qquad x - x\theta = n\theta - x\theta: \qquad \theta = x/n.$$

Prior: similarly, $f(\theta)$ has a maximum where $\log f(\theta)$ does, i.e. where

$$\frac{p-1}{\theta} - \frac{q-1}{1-\theta} = 0: \quad p - p\theta - 1 + \theta = q\theta - \theta: \quad \theta = (p-1)/(p+q-2).$$

Example 2. Normal family with normal prior ([O'H], Ex. 1.5 p.7). Suppose x is the sample mean of a sample of n independent readings from a normal

distribution $N(\theta, \sigma^2)$, with σ known and θ the parameter of interest. So x is $N(\theta, \sigma^2/n)$:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}.\sigma/\sqrt{n}} \exp\{-\frac{1}{2}(x-\theta)^2/\frac{\sigma^2}{n}\}.$$

Suppose that on the basis of past experience [prior knowledge] the prior distribution of θ is taken to be $N(\mu, \tau^2)$:

$$f(\theta) = \frac{1}{\sqrt{2}\pi\tau} \exp\{-\frac{1}{2}(\theta-\mu)^2/\tau^2\}.$$

Now $f(x) = \int f(\theta) f(x|\theta) d\theta$:

$$f(\theta) f(x|\theta) = \frac{1}{2\pi.\tau\sigma/\sqrt{n}}. \exp\{-\frac{1}{2}\Big[\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2/n}\Big]\}.$$

The RHS has the functional form of a bivariate normal distribution (IV.2 D7, [BF] 1.5). So to evaluate the θ-integration, we need to *complete the square* (cf. solving quadratic equations!). First,

$$(x-\theta)^2 = [(x-\mu)-(\theta-\mu)]^2 = (x-\mu)^2 - 2(x-\mu)(\theta-\mu) + (\theta-\mu)^2.$$

We write for convenience

$$c := \frac{1}{\tau^2} + \frac{1}{\sigma^2/n} :$$

$$\begin{aligned} f(\theta)f(x|\theta) &= const. \exp\{-\frac{1}{2}\Big[c(\theta-\mu)^2 - \frac{2}{\sigma^2/n}(\theta-\mu)(x-\mu) + \text{function of } x\Big]\} \\ &= const. \exp\{-\frac{1}{2}c\Big[(\theta-\mu)^2 - \frac{2(\theta-\mu)(x-\mu)}{c\sigma^2/n} + \text{function of } x\Big]\} \\ &= const. \exp\{-\frac{1}{2}c\Big(\theta - \mu - \frac{x-\mu}{c\sigma^2/n}\Big)^2 + \text{function of } x\}. \end{aligned}$$

Then from (*), to get the posterior density $f(\theta|x)$ we have to take the product $f(\theta)f(x|\theta)$ above, and divide by $f(x)$ – a function of x *only* (θ has been integrated out to get it). So: the posterior density $f(\theta|x)$ is itself of the form above, as a function of θ (with a different constant and a different function of x – but these do not matter, as our interest is in θ).

We can now recognise the posterior $f(\theta|x)$ – it is *normal*. We can read

off:
(i) its mean, $\mu + (x - \mu)/(c\sigma^2/n)$,
(ii) its variance, $1/c$. Thus the *posterior precision* is c. But from the definition of c, this is the sum of $1/\tau^2$, the *prior precision*, and $1/(\sigma^2/n)$, the *data precision*. By (i), the mean is

$$\mu[1 - \frac{\text{data precision}}{\text{posterior precision}}] + x.[\frac{\text{data precision}}{\text{posterior precision}}],$$

or

$$\mu[\frac{\text{prior precision}}{\text{posterior precision}}] + x.[\frac{\text{data precision}}{\text{posterior precision}}].$$

This is a *weighted average* of the prior mean μ and the data value x (the sample mean of the n readings), *weighted according to their precisions*. So:
(a) the form, mean and variance (or precision) of the posterior density are intuitive, statistically meaningful and easy to interpret,
(b) the conclusions above show clearly how the Bayesian procedure synthesises prior and data information to give a compromise,
(c) the family of normal distributions is closed in the above example: a normal prior and normal data give a normal posterior. This is an example of *conjugate priors*, to which we return later.
Note. The example above on the normal distribution makes another important point: often θ will be a vector parameter, $\theta = (\theta_1, \cdots, \theta_p)$ – as with, e.g., the normal distribution $N(\mu, \sigma^2)$. For simplicity, the variance σ^2 in the above was taken known. But in general, we will not know σ^2. Instead, we should include it in the Bayesian analysis, representing our uncertainty about it in the prior density. We then arrive at a posterior density $f(\theta|x)$ for the vector parameter $\theta = (\theta_1, \cdots, \theta_p)$. If our interest is in, say, θ_1, we want the posterior density of θ_1, $f(\theta_1|x)$. We get this just as in classical statistics we get a marginal density out of a joint density – by *integrating out the unwanted variables*.

In the normal example above, Ex. 2, we could impose a prior density on σ without assuming it known. This can be done ([O'H], Ex. 1.6 p.8, Lee [L], §2.12), but there is no obvious natural choice, so we shall not do so here.

Example 3. The Dirichlet distribution ([O'H], Ex. 1.7 p.10, §10.2-6). Consider the density in $\theta = (\theta_1, \cdots, \theta_k)$ on the region

$$\theta_1, \cdots, \theta_k \geq 0, \qquad \theta_1 + \cdots + \theta_k = 1$$

(a *simplex* in k dimensions), with density

$$f(\theta) \propto \Pi_{i=1}^{k} \theta_i^{a_i - 1}$$

for constants a_i. We quote that the constant of proportionality is

$$\Gamma(a_1 + \cdots + a_k)/\Gamma(a_1) \cdots \Gamma(a_k),$$

by *Dirichlet's integral*, an extension of the Eulerian integral for the gamma function (see [O'H] 10.4, or, say, 12.5 of
WHITTAKER, E. T. & WATSON, G. N.: *Modern analysis*, 4th ed., 1927/1963, CUP).
Thus the *Dirichlet density* $D(a_1, \cdots, a_k)$ with *parameters* $\theta_1, \cdots, \theta_k$ is

$$f(\theta) := \frac{\Gamma(a_1 + \cdots + a_k)}{\Gamma(a_1) \cdots \Gamma(a_k)} . {\theta_1}^{a_1 - 1} \cdots {\theta_k}^{a_k - 1}.$$

Now draw a random sample of size n from a population of k distinct types of individuals, with proportions θ_i of type i ($i = 1 \cdots k$). Then the likelihood is

$$f(x|\theta) = \frac{n!}{x_1! \cdots x_k!} . {\theta_1}^{x_1} \cdots {\theta_k}^{x_k},$$

the multinomial distribution. So

$$f(x|\theta) f(\theta) = const. {\theta_1}^{x_1 + a_1 - 1} \cdots {\theta_k}^{x_k + a_k - 1},$$

and the posterior density $f(\theta|x)$ is also of this form, with a different constant (making it a density - i.e., integrating to 1). We recognise the functional form of a Dirichlet density, with a_i replaced by $a_i + x_i$. So

$$f(\theta|x) = \frac{\Gamma(a_1 + \cdots + a_k + n)}{\Gamma(a_1 + x_1) \cdots \Gamma(a_k + x_k)} . {\theta_1}^{a_1 + x_1 - 1} \cdots {\theta_k}^{a_k + x_k - 1}$$

(as $x_1 + \cdots x_k = n$, the sample size).
Example 4. Poisson and Gamma distributions ([O'H], Ex. 1.1, 1.2 p.21).

Data: $x = (x_1, \cdots, x_n)$, x_i independent, Poisson distributed with parameter θ:

$$f(x|\theta) = \Pi_1^n f(x_i|\theta) = \theta^{x_1 + \cdots + x_n} e^{-n\theta} / x_1! \cdots x_n! = \theta^{n\bar{x}} e^{-n\theta} / \Pi x_i!,$$

where $\bar{x} := \frac{1}{n}\Sigma x_i$ is the sample mean.
Prior: the Gamma density $\Gamma(a,b) \quad (a,b>0)$:

$$f(\theta) = \frac{a^b \theta^{b-1}}{\Gamma(b)} e^{-a\theta} \qquad (\theta > 0):$$

$$f(x|\theta) f(\theta) = \frac{a^b}{\Gamma(b)\Pi x_i!} \theta^{n\bar{x}+b-1} e^{-(n+a)\theta},$$

$$f(\theta|x) \propto f(x|\theta) f(\theta) = const.\theta^{n\bar{x}+b-1} e^{-(n+a)\theta}.$$

This has the form of a Gamma density. So, it *is* a Gamma density, $\Gamma(n+a, n\bar{x}+b)$:

$$f(\theta|x) = \frac{(n+a)^{n\bar{x}+b}}{\Gamma(n\bar{x}+b)} . \theta^{n\bar{x}+b-1} e^{-(n+a)\theta} \qquad (\theta > 0).$$

Means. For $\Gamma(a,b)$, the mean is

$$E\theta = \int_0^\infty \theta f(\theta) d\theta = \frac{a^b}{\Gamma(b)} . \int_0^\infty \theta^b e^{-a\theta} d\theta.$$

Substituting $t := a\theta$, the integral is $\Gamma(b+1)/a^{b+1}$, which is $b\Gamma(b)/a^{b+1}$ using the functional equation for the Gamma function, $\Gamma(x+1) = x\Gamma(x)$. So the mean is $E\theta = b/a$. Similarly,

$$E\theta^2 = \int_0^\infty \theta^2 f(\theta) d\theta = \Gamma(b+2)/a^{b+2},$$

so $var\theta = E(\theta^2) - [E\theta]^2 = b(b+1)/a^2 - (b/a)^2 = b/a^2$.

So by above, the prior mean is b/a; the posterior mean is $(n\bar{x}+b)/(n+a)$; the data mean is $\bar{x}$. Write

$$\lambda := a/(n+a), \qquad \text{so } 1-\lambda = n/(n+a): \qquad \text{since}$$

$$\frac{n\bar{x}+b}{n+a} = \frac{a}{n+a} . \frac{b}{a} + \frac{n}{n+a} . \bar{x},$$

posterior mean $(n\bar{x}+b)/(n+a) = \lambda$. prior mean $b/a + (1-\lambda)$. sample mean $\bar{x}$.

Again, this is a weighted average, with weights proportional to n and a. Now n is the sample size, a measure of the precision of the data, and a is the rate of decay of the Gamma density, a measure of the precision of the prior information.

5. Pros and cons

5a. Advantages of the Bayesian paradigm

1. *Updating.*

Bayesian procedures provide an efficient algorithm for updating prior information as new data information is obtained. This is attractive theoretically: it reflects the way we all constantly update our thinking in the light of new experience, and it works well in a range of examples, as VII.4 shows. It also works well in many practical situations. It is particularly well suited to situations involving *time*, when new information is constantly coming in. Recursive algorithms exist for handling such situations *on-line*, or *in real time*, using computers. Such algorithms are typically Bayesian; an example is the *Kalman filter* (V.11 D9), used for on-line control problems (e.g., adjusting orbits of satellites) from the 1960s on.

2. *Uncertainty.*

We have seldom used the words 'probability' or 'random' in the above. Technically, Bayesian statistics differs from classical statistics by treating parameters, not as unknown constants, but – in effect (and explicitly, in [O'H]) – as random variables. This is necessary: only random variables can have distributions, prior and/or posterior.

This change of view – away from thinking of random variables and parameters as separate, towards treating them on the same footing, thinking about uncertainty – is often helpful, *provided* one takes the trouble to get used to it. This chapter is designed to do just that!

Some Bayesians carry this shift away from probability language to surprising extremes. An example is the famous dictum by the father of 20th century Bayesian statistics, Bruno de FINETTI (1906-1985):

PROBABILITY DOES NOT EXIST!

We would not go so far, but do recommend the Bayesian viewpoint as being useful and workable.

3. *Subjectivity.*

The information in the data is objective: it is the same to all statisticians following the same procedure and obtaining that data. By contrast, the information used in the choice of prior is subjective: it reflects the experience/knowledge/beliefs of the statistician (or his client). This subjectivity persists into the posterior distribution after we use Bayes' Theorem: the entire analysis has been *personalised*, to suit the statistician (or his client).

4. *Decision Theory.*

The Bayesian formulation (or paradigm) combines well with the ideas of Decision Theory. For this important subject, see e.g. [L].

One context in which the Bayesian/decision-theoretic approach is useful is in business/finance/investment. Suppose one is faced with the need to take major business decisions – e.g., whether/where/when to drill for oil. Drilling is very expensive, and may well produce no return on the large investment of capital in the shape of exploitable oil reserves. But, commercially viable oil reserves can be profitably exploited – and necessarily have to be found by risky exploratory drilling. Nothing venture, nothing win. This area involves *real options*, or *investment options*; see e.g. [Math428], VI.6 Week 11.

In such situations, the Bayesian approach quantifies one's uncertainty: decision theory then helps one to act rationally given one's beliefs.

5. *Output.*

The end-product of a Bayesian analysis is a *posterior distribution*. This is more informative than

(i) a number [point-estimate: e.g., a maximum-likelihood estimate],

(ii) two numbers [interval estimate: e.g., a confidence interval].

It also depends continuously on what it depends on – the prior information and the data information. The discontinuous 'accept or reject' framework of hypothesis testing is avoided.

6. *Nuisance parameters.*

A *nuisance parameter* is what its name implies: a parameter which is present in the formulation of the model, but absent from the question of interest. The parameter(s) in which we are interested are called, by contrast, *parameters of interest* or *interest parameters*.

E.g.: *Testing for equality of two normal means.*

The usual classical assumption for testing $H_0 : \mu_1 = \mu_2$ v. $H_1 : \mu_1 \neq \mu_2$, for two normal populations $N(\mu_i, \sigma_i^2)$, is to assume *equality of variances*: $\sigma_1 = \sigma_2$. Testing for equality of means *without* assuming equality of variances is a famous statistical problem, the *Behrens-Fisher problem*. It has a satisfactory solution (Scheffé's solution) when the two sample sizes n_1, n_2 are equal, but not in general.

E. g.: *Testing for normality.* Is this population normal? Here *both* μ *and* σ are nuisance parameters. It is much easier to ask: is this population $N(\mu_0, \sigma_0)$ for *specified* μ_0, σ_0? than to ask: is it $N(\mu, \sigma)$ for *some* μ, σ? One approach would be to estimate the mean and variance from the data, and then 'plug in' these estimates to try to reduce the second question to the first – but this

sort of procedure can be hard to justify.

In principle, nuisance parameters are easily handled in Bayesian statistics. If $\theta = (\theta_1, \theta_2)$ with θ_1 the interest parameter and θ_2 the nuisance parameter (either or both of θ_1, θ_2 can be several-dimensional), one finds the posterior density $f(\theta|x)$ as usual. This is the *joint* density of θ_1 and θ_2 (given the data x), so one extracts the *marginal* density of θ_1 (given x) as usual, by integrating out the unwanted variable θ_2:

$$f(\theta_1|x) = \int_{-\infty}^{\infty} f(\theta|x)d\theta_2 = \int_{-\infty}^{\infty} f(\theta_1, \theta_2|x)d\theta_2.$$

Of course, the integration may be difficult to perform – it may, in practice, need to be done numerically. But such problems are quite general, and not the fault of Bayesian statistics!

7. *The Likelihood Principle.*

As the fundamental formula of Bayesian statistics,

posterior density is proportional to prior density times likelihood

shows, the data only enters a Bayesian analysis through the likelihood. The *Likelihood Principle* (LP), formulated by G. A. BARNARD (1915-2002) (in a series of papers, 1947-1962) and A. BIRNBAUM (1962) says that the data should only enter any statistical analysis through the likelihood. Thus

Bayesian statistics satisfies the Likelihood Principle.

Classical statistics, however, violates the LP. O'Hagan, for instance, discusses a number of examples, including ([O'H] 33):
Bernoulli trials, success probability θ. Consider two situations:
(a) n trials; you observe r successes;
(b) toss till you observe the rth success: you need n trials.
The two likelihoods are the same [apart from constant factors, arising because in (b), but not in (a), the last toss must be a success]: to a Bayesian statistician, these situations are equivalent. To a classical statistician, however, they are quite different. For instance, the stopping rules are quite different [the area of statistics where one continues sampling until something happens and then stops is called Sequential Analysis]. [O'H] (5.14-15) points out that
(a) the minimum variance unbiased estimators of θ differ in these two cases;

(b) the very concept of unbiasedness itself violates the LP. For, it involves an expectation over the distribution of x - the bias in a statistic $T(x)$ is

$$b := E[T(x)|\theta] - \theta$$

– and this involves values of x we could have seen but didn't. The LP insists we take account only of the values of x we did see.

For a full-length (pro-Bayesian) account of the LP, see
BERGER, J. O. & WOLPERT, R. L. (1988): *The Likelihood Principle* (2nd ed.), Institute of Mathematical Statistics.

8. MCMC and the Gibbs sampler.

Modern computing power has made many previously intractable Bayesian implementations possible in practice. Key theoretical advances here are MCMC (Markov Chain Monte Carlo) and the Gibbs sampler (Gelfand & Smith 1990).

5b. Disadvantages of the Bayesian paradigm.

1. *Choice of prior.*

A Bayesian analysis cannot even begin without a choice of prior density (or distribution). This may well be problematic:
(a) we may have little prior information;
(b) what prior information we have may not suggest a mathematically convenient, or even tractable, choice of prior;
(c) the choice may be to some extent arbitrary;
(d) different choices of prior may (will) lead to different conclusions;
(e) we may have too sparse a collection of suitable families of priors to hand.
Of course, problems of this sort affect classical parametric statistics too. But classical statistics can fall back in such cases on a non-parametric approach, for which Bayesian treatments are less well developed, and in any case the problem is more acute in Bayesian statistics, as we have to choose suitable forms for both the prior and the likelihood.
Undoubtedly, choice of prior is the hardest thing in many – or even most – Bayesian analyses, and is the feature of Bayesian statistics most objectionable to non-Bayesians.

2. *Prior ignorance.*

The less a Bayesian knows, the harder he finds it to choose a prior. The worst-case scenario for a Bayesian is little (or even no) prior knowledge. To a

non-Bayesian, this is a non-problem: simply use a classical analysis, relying on the data (which is all we've got).

If θ belongs to a finite interval, $[a, b]$ say, there is a natural choice of prior to represent prior ignorance: the uniform density on $[a, b]$:

$$f(\theta) := 1/(b-a) \qquad \text{if } a \leq \theta \leq b, \text{ 0 else.}$$

But, there is no analogous density in an infinite interval - the real line, say. If $f(\theta) \equiv c > 0$, then *either* $c = 0$, when $\int_{-\infty}^{\infty} f(\theta)d\theta = 0$, *or* $c > 0$, when $\int_{-\infty}^{\infty} f(\theta)d\theta = +\infty$. It is impossible to get $\int_{-\infty}^{\infty} f(\theta)d\theta = 1$, the condition for $f(\theta) \geq 0$ to be a density, without $f(\theta)$ varying with θ. But this treates some θ-values differently from others, which is inconsistent with prior ignorance, when we have no grounds to discriminate between different values of θ.
Note. Some Bayesian statisticians have advocated using *improper priors* (allowing $\int_{-\infty}^{\infty} f(\theta)d\theta = +\infty$) in such cases, for this reason. But this is hard to justify, and is becoming less common nowadays.
3. *Objectivity.*

The Bayesian paradigm is well suited to situations where a subjective view is appropriate – particularly where a decision-taker has to act in the face of uncertainty, as in Decision Theory. Typical examples include businessmen facing management decisions about investment (whether/where/when to drill for oil, for instance). The manager's judgement is fed into the choice of prior, and he stands or falls by it. The subjective view is appropriate here.

By contrast, in science, one seeks objectivity. Whether or not Nature works in a certain way depends on Nature (or God), not on our opinions or beliefs [we leave to one side foundational questions about quantum mechanics, and whether or not a quantum formulation necessarily involves the mind of the observer]. Consequently, the Bayesian paradigm has met with more resistance in science than in business, because of the higher value put there on objectivity as against subjectivity.
Note. Lee's book makes telling use of examples about dating rocks in geology. Obviously the age of a rock (some hundreds of millions of years old) is completely objective – it hs nothing to do with us or our opinions. Indeed, it is hard to imagine anything more indifferent to us than a chunk of rock. It has a definite age; God (or Nature) knows this, but won't tell us. We thus have no means, even in principle, of assessing the age of a rock sample (which long predates humanity!) other than our own experimentation, observation and analysis, which will provide partial knowledge with remaining

uncertainty. The Bayesian paradigm does provide a sensible way of expressing this. So, despite the obvious objection about subjectivity, a Bayesian approach is quite defensible where, as here, it produces sensible results and there is nothing else to do.

4. *Summary statistics and dimensionality.*

For a one-dimensional parameter θ, the output is a posterior density, which we can graph. This is an advantage: 'One picture is worth a thousand words'! The advantage is particularly telling if, as we assume, a computer graphics capability is available. For a two-dimensional parameter θ, the output is a posterior density in the plane, which we can 'graph' in three dimensions, using a suitable computer graphics package. Again, this is an advantage. In *three* dimensions, graphics are no longer applicable, because *four* dimensions would be needed.

In higher dimensions, the situation rapidly gets even worse. We cannot *graph* the output; it becomes increasingly difficult even to *visualise* the output. Instead, we seek to *summarise* the output, using suitable summary statistics (e.g., mean/median/mode, covariance matrix, measures of skewness/kurtosis, ...). Thus the extra information in the Bayesian output (posterior density), over and above that from a classical output (summary statistics), is no longer an advantage – because we cannot use it – but actually a drawback – because we have to work to get back to summary statistics, such as a classical treatment provides anyway.

Note. 1. Summarisation methods are discussed in detail in [O'H], 2.1 – 2.24.

2. The dimensional aspects above underscore the *principle of parsimony*: one should seek to work in as low a dimensionality (i.e., with as few parameters) as possible. [It is quite common to find the complexity of a theory growing uncontrollably with increase in dimension. This phenomenon is called the *curse of dimensionality*, a term due to Richard Bellman.]

3. If the right dimensionality is not clear, we may be able to formalise the trade-off between the better fit a higher dimension can provide against the extra complexity by using methods such as Akaike's Information Criterion (AIC): see e.g. [BF] 5.2.1, [O'H], Ch. 7.

5. *Integration.*

Bayesian statistics involves the need to integrate in several ways:

(i) to get $f(x|\theta)f(\theta)d\theta$,

(ii) to get marginal posterior densities from joint posterior densities - e.g., to eliminate nuisance parameters,

(iii) to produce summary statistics as above - e.g., posterior means, etc.
Such integrations may be hard or impossible to do analytically. We may need to integrate numerically. This may be computer-intensive, and involves a good knowledge of, e.g.,
(a) numerical analysis as a branch of mathematics,
(b) computer implementation - e.g., by using the NAG Library [NAG = Numerical Algorithms Group, Oxford University].

Since c. 1990 (e.g. the Gelfand-Smith paper in *JASA*), much theoretical and practical progress in such areas has been made, using *Markov Chain Monte Carlo* (MCMC) methods – techniques such as the *Metropolis-Hastings algorithm* (VI.4 above and VII.6 below) and the *Gibbs sampler*. Such methods are extensively used nowadays (see e.g. the MSc in Statistics here).

6. Hierarchical models; Markov Chain Monte Carlo (MCMC).

In the Bayesian paradigm, everything is random, including the parameters; also, the parameters are drawn from a prior, and we may have difficulty in choosing the prior. Such difficulties may be lessened if we draw the parameters of the prior from some 'prior prior', which will itself have parameters, called *hyperparameters*. Such a model is called a *hierarchical model*. Our main sources here are Robert [R] Ch. 8,9, Gelman et al. [GCSR] Ch. 5, 11.

Definition. A *hierarchical Bayes model* is a Bayesian model $(f(x|\theta), \pi(\theta))$ in which the prior $\pi(\theta)$ is decomposed into conditional distributions

$$\pi_1(\theta|\theta_1), \pi_2(\theta_1|\theta_2), \ldots, \pi_n(\theta_{n-1}|\theta_n)$$

and a marginal $\pi_{n+1}(\theta_n|\theta_n)$ such that

$$\pi(\theta) = \int \ldots \int \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2)\pi_n(\theta_{n-1}|\theta_n)\pi_{n+1}(\theta_n)d\theta_1 \ldots d\theta_{n+1}. \qquad (H)$$

The parameters θ_i are called *hyperparameters of level* i.

A hierarchical Bayes model is itself a Bayesian model, but the decomposition (H) is often useful – e.g., in MCMC (below), and in revealing structural information.

One rarely needs to go beyond $n = 2$, and we shall not do so. So we shall always have

$$\theta|\theta_1 \sim \pi_1(\theta|\theta_1), \qquad \theta_1 \sim \pi_2(\theta_1). \qquad (H)$$

Here the distribution of θ is a *mixture* of the θ_1, with *mixing distribution* π_2.

Example: Random effects in the linear model.

We may have a *mixed model*, with some *fixed effects*, as in IV, and some *random effects*. The classical instance of this is Henderson's work on the breeding of dairy cows (1950). The fixed effects are the objects of study – typically, diet, of interest for its effect on milk yield. The random effects are the animals – animals differ, just as people do. It is conventional to write the model equation here as

$$y = X\beta + Zu + \epsilon,$$

where

$$W = (X, Z)$$

is the $n \times (p + q)$ design matrix, X $(n \times p)$ and Z $(n \times q)$ are the design submatrices for the fixed and random effects. We take the random effects u and the error ϵ uncorrelated (independent when both are Gaussian, as we may as well assume here). The best linear unbiased estimator (BLUE) of IV.1 is conventionally called a *best linear unbiased predictor (BLUP)* here. These are the solutions of *Henderson's mixed model equations (MMEs)*. Two different forms of the BLUP are given in [BF] 9.1. The use of Bayes' theorem is mentioned there. This is a hierarchical model with

$$y|\theta \sim N(\theta, \Sigma_1), \qquad \theta|\beta \sim N(X\beta, \Sigma_2).$$

Here the mean θ of y is decomposed into the fixed effects $X\beta$ and the random effects $Z\eta$, where $\eta \sim N(0, \Sigma_2)$.

Education.

Mixed models are widely used in educational studies (and more widely in Social Statistics). Here the fixed effects are the ones being studied – concerning, e.g., influence on performance of changes in syllabus, examination mode etc. The random effects are the pupils.

Finance.

Here the fixed effects are state of the economy, industrial sector etc. The random effects are the specific characteristics of the individual firms involved in the study.

Bayesian v. classical.

Strictly speaking, whether this procedure is classical or Bayesian depends

on what our inference is about. The procedure is classical if the inference is about the fixed effects (β), but Bayesian if it is about the overall effects (θ).

Normal mean-variance mixtures (NMVM); normal variance mixtures (NVM).

The *Bessel function of the third kind*, or *Macdonald function*, K_λ (λ real) is defined (for our purposes) by the integral representation

$$K_\lambda(x) = \frac{1}{2}\int_0^\infty u^\lambda \exp\{-\frac{1}{2}(u + 1/u)\}du/u \qquad (x \geq 0).$$

Then for $\psi, \chi > 0$,

$$f(x) := \frac{(\psi/\chi)^{\frac{1}{2}\lambda}}{2K_\lambda(\sqrt{\psi\chi})} x^{\lambda-1} \exp\{-\frac{1}{2}(\psi x + \chi/x)\} \qquad (x > 0)$$

is a probability density, the *generalised inverse Gaussian (GIG).*

The distribution of $x \sim N(\mu + \beta\sigma^2, \sigma^2)$, where σ^2 is sampled randomly from GIG, forms a *normal mean-variance mixture (NMVM)*, with *mixing distribution GIG*. It is called the *generalised hyperbolic* distribution, GH. The case $\beta = 0$ is simpler; we then get a *normal variance mixture (NVM).*

The GH distributions have been much used in mathematical finance, specially for return distributions with intermediate return interval – say, daily returns (Bingham & Kiesel 2001; Barndorff-Nielsen 1970s-90s; Eberlein 1990s). The log-density is a (branch of a) hyperbola (hence the name). As a hyperbola has linear asymptotes, the log-density decays linearly at $\pm\infty$. By contrast, the Gaussian log-density (monthly returns) decays quadratically, while the Student t log-density (tick data) decays logarithmically.

The GH distributions can be defined in any number of dimensions. They have two important general properties:
1. They are *elliptical*. They are an important parametric special case within this semi-parametric setting; see I.6.2 D2, V.6 D6, VI.3 D10.
2. They are *self-decomposable*: they belong to the class SD of distributions of stationary $AR(1)$ time-series models,

$$X_t = \rho X_{t-1} + \epsilon_t.$$

Bayesian sampling; HM.

We return to (H), in the form

$$\pi(\theta|x) = \int \pi_1(\theta|x, \lambda)\pi_2(\lambda|x)d\lambda. \qquad (H)$$

If we can sample efficiently from π_1 and π_2, we can use MCMC (in the form of a Bayesian sampling technique, *data augmentation* (Tanner & Wong, 1987)) to sample from π, by the following iterative algorithm.
Initialisation: Start with an arbitrary value λ_0.
Iteration: For $i = 1, \ldots, k$, generate
a. $\theta_i \sim \pi_1(\theta|x, \lambda_{i-1})$;
b. $\lambda_i \sim \pi_2(\lambda|x, \theta_i)$.
The generation of θ_i only depends on θ_{i-1}, not on previous values, so (θ_i) has the Markov property. Under suitable regularity conditions, this Markov chain will be ergodic, with limiting distribution π; furthermore, the approach to stationarity will often be geometrically fast.

The Hastings-Metropolis algorithm HM in this setting runs as follows. To sample from a distribution π known up to a normalising factor, and given a transition kernel $q(\theta|\theta')$, HM proceeds as follows.
(i) Start with θ_0 arbitrary.
(ii) Update from θ_m to θ_{m+1} by:
1. Generate $\xi \sim q(.|\theta_m)$;
2. Define

$$\rho := \Big(\frac{\pi(\xi) q(\theta_m|\xi)}{\pi(\theta_m) q(\xi|\theta_m)} \Big) \wedge 1.$$

3. Take

$$\theta_{m+1} := \xi \quad \text{with probability } \rho, \quad \theta_m \quad \text{otherwise.}$$

Again under suitable regularity conditions, the Markov chain (θ_m) converges to the equilibrium distribution π as m increases. The convergence is often geometrically fast, again under suitable conditions.

Graphical models

It is possible to model complex statistical situations, with many variables, some of which are *conditionally independent given others*. Such conditional independence can be conveniently encoded, and represented visually, using *graphs* (in the sense of Graph Theory, an important branch of Combinatorial Theory). We must be brief here; we refer for a monograph treatment to
Steffen L. LAURITZEN, *Graphical models*, OUP, 1996.

Graphical models originate in three different areas:
(i) Statistical Physics, in the work of Gibbs[1]. Here the idea is that particles

[1]J. W. Gibbs (1839-1903), American; one of the three founding fathers of Statistical Physics, with James Clerk Maxwell (1831-1879), Scottish, and Ludwig Boltzmann (1844-

can only interact with their immediate neighbours.
(ii) Genetics. This, incidentally, is one of the major application areas of heirarchical models, MCMC etc. (Human Genome Project, etc.).
(iii) Contingency tables. The analysis of complicated multi-dimensional contingency tables, where the data is counts cross-classified by characteristics, is important in the Social Sciences.

See in particular Lauritzen, Ch. 4 (Contingency tables), Ch. 5 (Multivariate normal models), 7.3.1 (MCMC); also *EM algorithm* (two steps – expectation, maximisation), 7.4.1.

Postscript.
1. Bayesian solution of the equity premium puzzle.

Following Markowitz (I.5), we should diversify our financial savings into a range of assets in our portfolio – including cash (invested risklessly – e.g., by buying Government bonds, or 'gilts', or putting it in the bank or building society – which we suppose riskless here, discounting such disasters as the Icelandic banking crisis, Northern Rock, RBS etc.) and risky stock. There is no point in taking risk unless we are paid for it, so there will be an excess return – equity premium – $\mu - r$ of the risky stock (return μ) over the riskless cash (return r), to be compared with the volatility σ of the risky stock via the *Sharpe ratio* (or *market price of risk*) $\lambda := (\mu - r)/\sigma$). Historical data show that the observed excess return seems difficult to explain.

A Bayesian solution to this 'equity premium puzzle' (the term is due to Mehra & Prescott (1985)) has been put forward by Jobert, Platania and Rogers: there is no equity premium puzzle, if one uses a Bayesian analysis to reflect fully one's uncertainty in modelling this situation. See
[JPR] A. JOBERT, A. PLATANIA & L. C. G. ROGERS, *A Bayesian solution to the equity premium puzzle*. Preprint, Cambridge (available from Chris Rogers' homepage: Cambridge University, Statistical Laboratory).
The Twenties Example [JPR]. One observes daily prices of a stock for T years, with an annual return rate of 20% and an annual volatility of 20%. How large must T be to give confidence intervals of $\pm 1\%$ for (i) the volatility, (ii) the mean? Answers: (i) about 11; (ii) about 1,550!!

This illustrates what is called *mean blur*; see e.g.
D. G. LUENBERGER, *Investment Science*, OUP, 1997.
Rough explanation: for the mean, only the starting and final values matter

1906), German.

(so effective sample size is 2); for the volatility, everything matters.

For non-Bayesian approaches, see e.g. Maenhout, *Rev. Fin. Studies* (2004), Barillas, Hansen & Sargent, *J. Econ. Th.* (2009).

2. Bayesian Non-parametrics.

We have discussed Bayesian statistics at some length in this Ch. VII, and (more briefly) Non-parametric statistics in Ch. VI. It is natural to wonder whether the two can be combined – and indeed, they can. This has been enormously helped by the growth of modern computer power. For those interested: e.g., Googling "Bayesian non-parametrics" produced 7,990 hits and "Bayesian nonparametrics" 30,700. There are lots of connections with machine learning, for example, and lots of applications. NHB