Bayesian Methods: Theory and Practice
Lecture 1 – Foundations

Harrison B. Prosper

Florida State University

CMS Statistics Committee

6 August, 2008

Many Thanks!

The lectures benefited from several useful comments from my colleagues on the CMS Statistics Committee, in particular, from Bob Cousins and Luc Demortier.

A Short (!) Reading List

- **Books**
  - **A. O'Hagan**, *Kendall's Advanced Theory of Statistics*, Volume 2B: *Bayesian Inference*, Oxford University Press (1994)
  - L. J. Savage, *The Foundations of Statistics*, Wiley (1954)
  - **D.S. Sivia and J. Skilling**, *Data analysis: A Bayesian Tutorial*, 2nd ed., Oxford University Press (2006)
  - H. Jeffreys, *Theory of Probability*, 3rd edition, Oxford University Press (1961)
  - G.E.P. Box and G.C. Tiao, *Bayesian Inference In Statistical Analysis*, John Wiley & Sons (1992)
  - E.T. Jaynes and L. Bretthorst, *Probability Theory, the Logic of Science*, Cambridge University Press (2003); http://omega.math.albany.edu:8008/JaynesBook.html

  - **S.K. Chatterjee**, *Statistical Thought: A Perspective and History*, Oxford University Press (2003)

Blue – A good starting point

A Short (!) Reading List

- **Articles**
  - R.T. Cox, *Probability, Frequency, and Reasonable Expectation*, Am. J. Phys. **14**, 1 (1946)
  - D. Heath and W. Sudderth, *de Finnetti's Theorem on Exchangeable Variables*, Am. Stat. 30 (4), 188 (1976)
  - R.E. Kass and L. Wasserman, *The Selection of Prior Distributions by Formal Rules*, J. Amer. Statist. Assoc., **91**, 1343 (1996)
  - J. Bernardo, Select Recent Papers, http://www.isds.duke.edu/research/conferences/valencia/publications.html
  - J. Berger, http://www.stat.duke.edu/~berger/papers.html
  - L. Daston, *How Probability Came To Be Objective And Subjective*, Hist. Math. **21**, 330 (1994)
  - C.M. Caves, *Probabilities as betting odds and the Dutch book*, http://info.phys.unm.edu/~caves/reports/dutchbook.pdf
  - H.B. Prosper, *Small Signal Analysis in High-Energy Physics: A Bayesian Approach*, Phys. Rev. **D37**, 1153 (1988)
  - R.D. Cousins, *Why Isn't Every Physicist A Bayesian*?, Am. J. Phys. **63**, 398 (1995)
  - L. Demortier, *Objective Bayesian Upper Limits for Poisson Processes*, http://physics.rockefeller.edu/~luc/technical_reports/cdf5928_objective_bayes_ul.pdf (2005)

Points to Note

- These lectures are a follow-up to the broad introduction given by Bob Cousins some weeks ago. The goal of these lectures is to provide a distillation of some key ideas found in the short reading list.

- However, from time to time, I shall feel compelled to inject my own viewpoint!

- The symbol ☺, at the start of a comment, signals a point of view I share, which is not necessarily endorsed by my colleagues on the CMS Statistics Committee.

Outline

- Lecture 1 – **Foundations**
  - Historical Sketch
  - Probability
    - Degrees of Belief
    - Avoiding Sure Loss
    - Exchangeability
  - The Bayesian Approach
- Lecture 2 – **Foundations & Applications**
- Lecture 3 – **Applications**

Historical Sketch

Before the 1850s, it was taken as self-evident by most mathematicians and scientists (e.g., Bayes, Bernoulli, Pascal, Poisson, Laplace, Lagrange, Maxwell,...) that **chance** and **probability** were quite distinct notions:

1. **chance** is a demonstrable property of the world: it is the unpredictability of certain outcomes, such as getting a six in the throw of a dice, while

2. **probability** is a measure of the **degree of belief** in propositions whose truth cannot be decided conclusively, given the limited information at hand.

Historical Sketch

In the late 19th century, following the work of John Venn, George Boole and others, and continuing into the early 20th century with the groundbreaking work of Fisher and Neyman to name but two, the view of probability as the limit, in some sense, of a **relative frequency** came to dominate statistical thinking.

Then in the 2nd half of the 20th century, following pioneering work in the 1930s, 1940s and 1950s by Jeffreys, de Finetti, Savage and others, the interpretation of probability as a measure of **degree of belief** was revived.

Probability

Probability

In 1933, the Soviet mathematician Andrey Kolmogorov created an axiomatic theory of probability comprising three elements: a set **Ω**, its **subsets** (specifically, a σ-algebra**) and a **measure***, **P**, called **probability**, that satisfies the axioms:

$$1: P(A) \geq 0$$

$$2: P(\Omega) = 1$$

$$3: P(A_1 + A_2 + \ldots) = \sum_i P(A_i), \quad \text{if } A_i \cdot A_j = 0, \forall i \neq j$$

1903 – 1987

* A measure is a function that assigns a magnitude to a set.

** A set of subsets of Ω, their complement and their union.

http://www-groups.dcs.st-and.ac.uk/~history/Mathematicians/Kolmogorov.html

Degrees of Belief

Consider the **propositions**, labeled A and B:

A = Obama will win in 2012

B = Pallin will win in 2012

We assume the following is possible, at least in principle: to each proposition a number can be assigned that quantifies the degree to which the proposition is thought to be plausible. These numbers are usually referred to as **degrees of belief**.

A priori, there is no obvious reason why these numbers should have anything to do with probability. However,…

Degrees of Belief

…several different lines of reasoning conclude that degrees of belief, **b**, follow rules

1. $b(A) \geq 0$
2. $b(A) = 1$ if A is true
3. $b(A) = 0$ if A is false

Sum Rule

4. $b(A+B) = b(A) + b(B)$ if AB is false*

Product Rule

5. $b(AB) = b(A|B)\, b(B)$*

that are the *same* as those of probability. Therefore, we are at liberty to *interpret* probability as degree of belief, that is, to make the identification **P = b**.

***A+B = A or B, AB = A and B, A|B = A given that B is true**

The Cox Axioms

One line of reasoning is due to the physicist Richard Cox who, in 1946, proposed axioms akin to the following

Axiom 1 Degrees of belief can be ordered

Axiom 2 b(A) and b(~A) are related

Axiom 3 b(AB), b(A|B) and b(B) are related

Axiom 1 is of course an idealization. After all, it is doubtful that a real person given a large set of propositions would be able to order them *consistently* according to his or her perceived degree of belief in each.

The Cox Axioms

Since, by assumption, degrees of belief can be ordered, they can be represented by real numbers *b*. Cox, and others, showed that after a suitable rescaling of the degrees of belief, they obey the same rules as those of probability.

The theory is completed by incorporating the axioms of the **algebra of propositions**, **Boolean algebra**

$A+0 = A$	$A1 = A$
$A+\sim A = 1$	$A\sim A = 0$
$A+B = B + A$	$AB = BA$
$A+BC = (A+B)(A+C)$	$A(B+C) = AB+AC$

Bayes' Theorem

Note that **A**B = B**A**. Therefore, the product rule

$$P(\mathbf{A}B) = P(\mathbf{A}|B)\,P(B)$$

yields

$$P(B\mathbf{A}) = P(B|\mathbf{A})\,P(\mathbf{A})$$

But, since $P(\mathbf{A}B) = P(B\mathbf{A})$

Bayes' Theorem

$$P(B|\mathbf{A}) = P(\mathbf{A}|B)\,P(B) \,/\, P(\mathbf{A})$$

follows immediately.

Note also that setting B = ~A in the sum rule and using the fact that the proposition A + ~A is identically true yields

$$P(A) + P(\sim A) = 1$$

Avoiding Sure Loss

Another interesting set of arguments about probability were developed by Bruno de Finetti, who proposed the following **operational definition** of probability:

If you are ***rational*** and you are ***willing*** to forfeit an amount **X** in the hope you may gain an amount **Y**, then your degree of belief in proposition A, that is, the probability P(A) you assign to it, is ***defined*** to be

$$P(A) = X / Y$$

Net gain if A is true: $G = Y - X = [1 - P(A)]Y$

Net gain if A is false: $G = -X = -P(A)\,Y$

Note, since, by assumption, you are rational, $P(A) \leq 1$.

Avoiding Sure Loss

Example 1

Consider propositions A, B and A+B, where A and B are **mutually exclusive**, that is, AB is false:

A	Obama will win in 2012
B	Pallin will win in 2012
A + B	Obama or Pallin will win in 2012

You *assign* probabilities P(A), P(B) and P(A+B) and forfeit the amounts P(A) Y_A, P(B) Y_B and P(A+B)Y_{A+B}, respectively, in view of potential gains Y_A, Y_B and Y_{A+B}. Think of these as bets on the outcomes!

Avoiding Sure Loss

Example 1

The three possible outcomes and your three possible ***net*** gains are:

1. A is **true** and B is **false**, yielding a net gain for you of
$$G_1 = \mathbf{[1 - P(A)]Y_A - P(B)\,Y_B + [1 - P(A+B)]Y_{A+B}}$$

2. A is **false** and B is **true**, yielding a net gain for you of
$$G_2 = \mathbf{-P(A)Y_A + [1 - P(B)]\,Y_B + [1 - P(A+B)]Y_{A+B}}$$

3. A is **false** and B is **false**, yielding a net gain for you of
$$G_3 = \mathbf{-P(A)Y_A - P(B)\,Y_B - P(A+B)Y_{A+B}}$$

Avoiding Sure Loss

Example 1

Now suppose you have no control over the values of the gains Y_A, Y_B, Y_{A+B}. Then it is possible for an unscrupulous agent to chose them so as to guarantee you suffer a loss, ***whatever the actual outcome***!

This scam is called a Dutch book.

The scam can be thwarted, however, provided that you assign probabilities such that ***no*** solution, Y_A, Y_B, Y_{A+B}, exists to the **gain equations**

$$G_1 = [1 - P(A)]Y_A - P(B)\,Y_B + [1 - P(A+B)]Y_{A+B}$$
$$G_2 = -P(A)Y_A + [1 - P(B)]\,Y_B + [1 - P(A+B)]Y_{A+B}$$
$$G_3 = -P(A)Y_A - P(B)\,Y_B - P(A+B)Y_{A+B}$$

Avoiding Sure Loss

Example 1

The gain equations

$$G_1 = [1 - P(A)]Y_A - P(B)\,Y_B + [1 - P(A+B)]Y_{A+B}$$
$$G_2 = -P(A)Y_A + [1 - P(B)]\,Y_B + [1 - P(A+B)]Y_{A+B}$$
$$G_3 = -P(A)Y_A - P(B)\,Y_B - P(A+B)Y_{A+B},$$

which can be written as a matrix equation

$$G = \mathbf{P}\,Y,$$

will have no solutions if the determinant, $|\mathbf{P}|$, of the matrix $\mathbf{P}$ is forced to be zero. The condition $|\mathbf{P}| = 0$, yields

$$P(A+B) = P(A) + P(B).$$

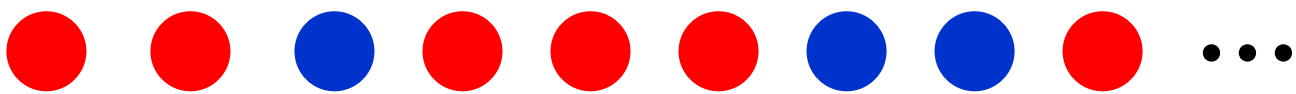
Analogous reasoning, yields the other probability rules.

Exchangeability

Consider a sequence of trials, each with only two possible outcomes: **success** or **failure**. Such trials are called **Bernoulli trials**.

A near perfect example of such trials are the ones we are about to start at the LHC: proton-proton collisions in which a success could be an event not described by the Standard Model.

Suppose we observe a sequence (***k***, ***n***) of ***k*** successes in ***n*** trials.



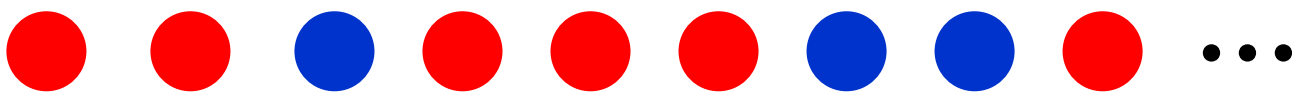
Exchangeability

What is the probability to get the sequence (***k***, ***n***)?

This cannot be answered in general. However, in the 1930s, de Finetti studied the consequences of imposing the following two conditions:

1. The order of any pair of trials can be **exchanged** without changing the probability of the sequence of outcomes.
2. The sequence (***k***, ***n***) can be embedded in a longer sequence (***r***, ***m***) whose length can be made arbitrarily long. The *unknown* **relative frequency** of success is $z = r / m$.

de Finetti called these two properties **exchangeability**.



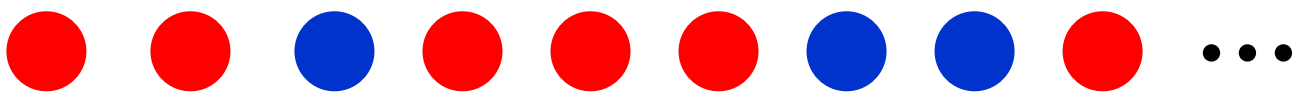
Exchangeability

The assumption of exchangeability forces all sequences (***k***, ***n***) to be assigned the same ***subjective*** probability $P(\boldsymbol{k}, \boldsymbol{n})$.

Exchangeability further implies that we can expand $P(\boldsymbol{k}, \boldsymbol{n})$ as follows

$$P(k,n) = \sum_{r=0}^{m} P(k,n \mid r,m) P(r,m)$$

for arbitrarily large ***m***, where $P(\boldsymbol{r}, \boldsymbol{m})$ is the ***subjective*** probability assigned to the longer sequence (***r***, ***m***), or, equivalently, to the *unknown* relative frequency $z = \boldsymbol{r} / \boldsymbol{m}$.



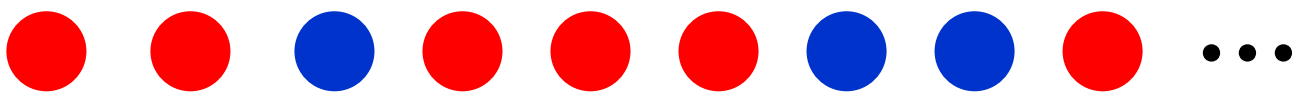
Exchangeability

$$P(k,n) = \sum_{r=0}^{m} P(k,n \mid r,m) P(r,m)$$

As $\boldsymbol{m}$ goes to infinity, the probability $P(\boldsymbol{k}, \boldsymbol{n}|\boldsymbol{r}, \boldsymbol{m}) \rightarrow \boldsymbol{z}^{\boldsymbol{k}}\ \boldsymbol{(1-z)}^{\boldsymbol{k}}$, where $\boldsymbol{z} = \boldsymbol{r} / \boldsymbol{m}$. Likewise, the probability $P(\boldsymbol{r}, \boldsymbol{m}) = P(\boldsymbol{zm}, \boldsymbol{m})$ coalesces into the **subjective prior density** $\pi(\boldsymbol{z})$.

Finally, when one accounts for the number of possible distinguishable sequences $(\boldsymbol{k}, \boldsymbol{n})$, one arrives at de Finetti's celebrated **representation theorem**

$$P(k \mid n) = \binom{n}{k} P(k,n) = \int_0^1 \text{Binomial}(k \mid n, z)\, \pi(z)\, dz$$



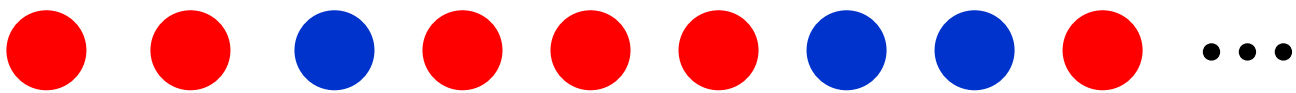
Exchangeability

If we know for *certain* that the relative frequency is $\boldsymbol{z} = \boldsymbol{p}$, or we have a *prediction* of it, then we can set

$$\pi(\boldsymbol{z}) = \delta(\boldsymbol{z} - \boldsymbol{p})$$

in which case the representation theorem reduces to the binomial distribution

$$P(\boldsymbol{k} \mid \boldsymbol{n}, \boldsymbol{p}) = \text{Binomial}(\boldsymbol{k} \mid \boldsymbol{n}, \boldsymbol{p})$$



Comments

Bayesian methods are based on the **subjective interpretation** of probability, that is, on the interpretation we have just sketched.

But Bayesians come in many flavors of which the most pungent are:

1. **Subjective Bayes**

and

2. **Objective Bayes**

Comments

Subjective Bayes

Every probability should be the *actual* degree of belief of a *real* person. If the degree of belief of a person can be elicited accurately this approach is **coherent** in the sense of **avoiding sure loss**. The coherence may not persist, however, when the inevitable short-cuts and approximations enter an analysis.

- ☺ But, coherence requires *only* that degrees of belief obey the rules of probability *strictly*.
- ☺ The stipulation that *all* degrees of belief be those of a *real* person does *not* follow from the probability rules. A better name for this approach is: **Personalistic Bayes** (Savage).

Comments

Objective Bayes

Degrees of belief can be the actual ones of real persons as well as those that approximate the degrees belief of an **ideal reasoner** whose knowledge, with respect to some aspects of the problem, is *minimal*. Since the ideal reasoner is an abstraction, elicitation of degrees of belief must be replaced by ***assessment using formal rules***.

However, the choice of rules is a matter of judgment. Moreover, they tend to yield degrees of belief that are *not* probabilities. Consequently, coherence has to be checked on a case by case basis.

Comments

☺ Probability provides a ***model*** of reasoning in the face of uncertainty. However, numerous psychological studies demonstrate that human beings follow a much more complex model that *appears*, in part, to be irrational*.

☺ Probability therefore should be regarded as a ***normative*** model: it tells us how we ***ought*** to reason when faced with uncertainty, not how we ***actually*** reason.

☺ I do not like the terms **subjective** and **objective** Bayes. The former is akin to speaking of wet water, while the latter is an oxymoron!

*See for example, *New Scientist*, Vol 199 No 2666, 26 July, 2008.

The Bayesian Approach

The Bayesian Approach

Definition:

A method is Bayesian if

1. it is based on probability interpreted as degrees of belief and
2. it uses Bayes' theorem

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

In practice, the form of Bayes' theorem most commonly used is based on **probability densities** that encode propositions about **data *D*** and **parameters ω**, which can be continuous, discrete, or both.

The Bayesian Approach

Following the statistician José Bernardo, we call

$p(D \mid \omega)$ the **probability model** that represents the mechanism that gave rise to the observed data D, given some *unknown* value of the parameters ω.

$\pi(\omega)$ the **prior probability density** over the parameter space Ω of the probability model and

$p(\omega \mid D)$ the **posterior density**.

As noted in Bob Cousin's talk, probabilities are *not* absolute, but are *always* conditional on context and assumptions.

The Bayesian Approach

The proximate goal of a Bayesian analysis is to compute the posterior density of the model parameters ω from Bayes' theorem*

$$p(\omega \mid D) = \frac{p(D \mid \omega)\pi(\omega)}{\int_{\Omega} p(D \mid \omega)\pi(\omega) d\omega}$$

Subsequently, one may use the posterior density to make **predictions**, compute **point** and **interval estimates** and perform **hypothesis tests**.

*In principle, in the personalistic approach, one has the option to elicit the posterior directly.

The Bayesian Approach

Example 2: Suppose one has observed D events. Let's *assume* that the mechanism that generated D can be modeled by

$$p(D \mid \omega) = \text{Poisson}(D|\omega) = \exp(-\omega)\ \omega^D / D!$$

The prior density $\pi(\omega)$ should reflect our prior beliefs about the Poisson parameter ω. Suppose we believe it to be confined to the interval $0 \leq \omega \leq \beta$ and that its value is closer to zero than to the upper bound. We might choose to model these prior beliefs using the prior density

$$\pi(\omega) \quad = \text{const.} / \omega^k, \text{ perhaps with } 0 \leq k < 1.$$

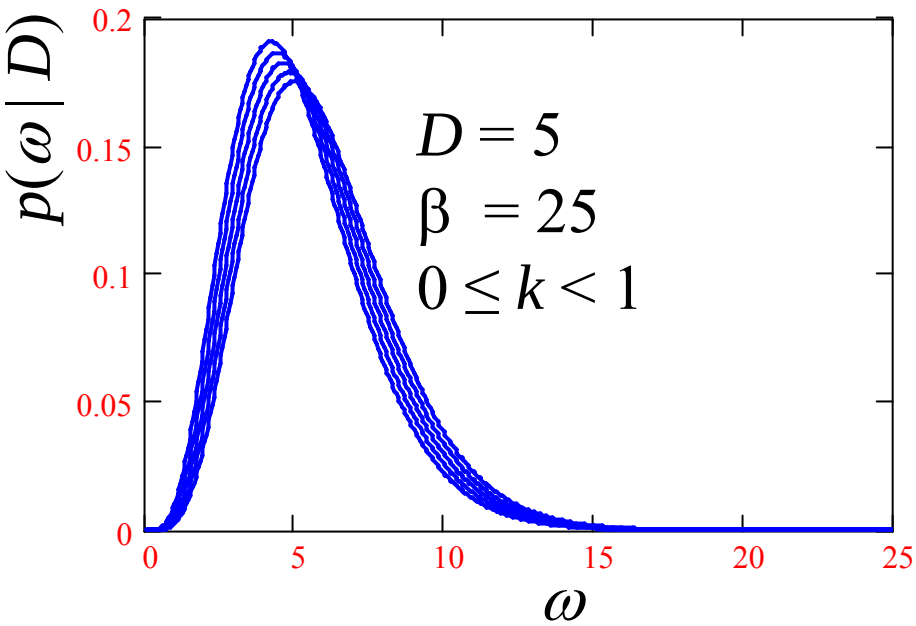
The Bayesian Approach

Example 2: The posterior $p(\omega \mid D)$ is readily calculated from Bayes' theorem

$$p(\omega \mid D) = \frac{\exp(-\omega)\omega^{D-k}}{\int_0^\beta \exp(-\omega)\omega^{D-k} d\omega}$$

$$= \frac{\exp(-\omega)\omega^{D-k}}{\gamma(D-k+1, \beta)}$$

where $\gamma(n, x)$ is the lower incomplete gamma function.



The plot shows $p(\omega \mid D)$ for different values of k.

The Bayesian Approach

Marginalization – If we are interested only in a subset θ of the parameters $\omega = \theta, \phi$, we need a way to restrict $p(\omega \mid D) = p(\theta, \phi \mid D)$ to $p(\theta \mid D)$.

Since the Bayesian approach is merely applied probability theory, the way to effect this restriction is specified, *uniquely*, by a *theorem* of probability theory: **marginalize** with respect to the remaining parameters

$$p(\theta \mid D) = \int p(\theta, \phi \mid D) d\phi$$

The remaining parameters ϕ are often referred to as **nuisance parameters**.

The Bayesian Approach

Prediction – Suppose we observe data D and we wish to predict the values of new data X. For example, in track fitting, given the hits on the current track, we may wish to determine where to look for potential new hits.

The solution, in principle, is to compute the **posterior predictive distribution**

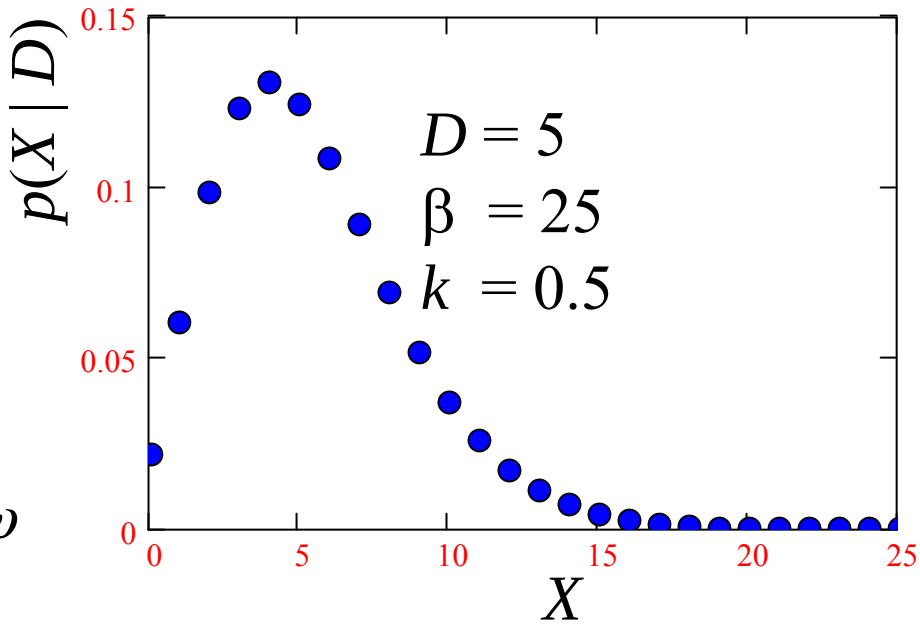
$$p(X \mid D) = \int p(X \mid \omega)\, p(\omega \mid D) d\omega$$

The **Kalman filter** and its variants, which are widely used in track fitting, can be thought of as linear approximations to the posterior predictive distribution.

The Bayesian Approach

Example 3: Given the posterior density $p(\omega \mid D)$ from Example 2

$$\frac{\exp(-\omega)\omega^{D-k}}{\gamma(D-k+1,\beta)}$$



its posterior predictive distribution is given by

$$p(X \mid D) = \int_0^\beta p(X \mid \omega)\, p(\omega \mid D)\, d\omega$$

$$= \frac{\gamma(X+D-k+1,\beta)}{2^{X+D-K+1}\,\gamma(D-k+1,\beta)\,X!}$$

Summary

- **Probability**
  - If we are prepared to make the idealization that degrees of belief can be represented by real numbers, one can establish through a variety of arguments that these numbers follow rules identical to those of probability.

- **Bayesian Approach**
  - Since this approach is based on the degree of belief interpretation of probability, it is ***irreducibly subjective***.
  - Opinion differs about whether the notion degree of belief should be restricted to real persons or whether one is permitted to extend it to ideal reasoners.

Summary

- **Bayesian Approach**
  - The proximate goal of a Bayesian analysis is to compute the posterior density of the model parameters.
  - A subsequent goal may include extracting summaries of the posterior density, such as point or interval estimates, using the posterior density to make predictions and/or to perform hypothesis tests.