

Module 1

A simple discrete example of the “Bayes paradigm”

Prof. Stephen B. Vardeman
Statistics and IMSE
Iowa State University

March 5, 2008

A Simple Discrete Example (Likelihood/Data Model)

The purpose of this module is to show precisely the mathematics of the “Bayes paradigm” in a computationally elementary context. Suppose that a random quantity X with possible values 1, 2, 3, and 4, has a distribution that depends upon some parameter θ that has possible values 1, 2, and 3. We’ll suppose the three possible distributions of X (appropriate under the three different values of θ) may be specified by probability mass functions $f(x|\theta)$ given in Table 1.

Table 1 Three Possible Distributions of X

x	$f(x 1)$
1	.4
2	.3
3	.2
4	.1

x	$f(x 2)$
1	.25
2	.25
3	.25
4	.25

x	$f(x 3)$
1	.1
2	.2
3	.3
4	.4

(“Big” parameter values will tend to produce “big” values of X .)

A Simple Discrete Example (Prior)

A “Bayes” approach to making inferences about θ based on an observation $X = x$ require specification of a “prior” probability distribution for the unknown parameter θ . There are myriad possibilities here. For sake of illustration, we will compare analyses based on distributions with probability mass functions $g_A(\theta)$, $g_B(\theta)$, and $g_C(\theta)$ given in Table 2.

Table 2 Three “Prior” Distributions for θ

θ	$g_A(\theta)$	θ	$g_B(\theta)$	θ	$g_C(\theta)$
1	.5	1	.3	1	.2
2	.3	2	.4	2	.3
3	.2	3	.3	3	.5

The first prior favors small parameter values, the second is “flatter” but slightly favors the “moderate” parameter value, and the third favors large parameter values.

A Simple Discrete Example (Joint)

A given prior distribution together with the forms specified for the distribution of X given the value of θ (the "likelihood") leads to a joint distribution for both X and θ with probability mass function $g(\theta, x)$ that can be represented in a two-way table, where any entry is obtained as

$$\begin{aligned}g(\theta, x) &= f(x|\theta) g(\theta) \\ &= \text{likelihood} \cdot \text{prior}\end{aligned}$$

For example, using the first prior distribution one obtains the joint distribution $g_A(\theta, x)$ specified in Table 3.

A Simple Discrete Example (Joint cont.)

Table 3 Joint Distribution for X and θ Corresponding to the First Prior for θ

	$\theta = 1$	$\theta = 2$	$\theta = 3$	
$x = 1$	$g_A(1, 1) = .4(.5) = .2$	$g_A(2, 1) = .25(.3) = .075$	$g_A(3, 1) = .1(.2) = .02$	$f_A(1) = .295$
$x = 2$	$g_A(1, 2) = .3(.5) = .15$	$g_A(2, 2) = .25(.3) = .075$	$g_A(3, 2) = .2(.2) = .04$	$f_A(2) = .265$
$x = 3$	$g_A(1, 3) = .2(.5) = .1$	$g_A(2, 3) = .25(.3) = .075$	$g_A(3, 3) = .3(.2) = .06$	$f_A(3) = .235$
$x = 4$	$g_A(1, 4) = .1(.5) = .05$	$g_A(2, 4) = .25(.3) = .075$	$g_A(3, 4) = .4(.2) = .08$	$f_A(4) = .205$
	$g_A(1) = .5$	$g_A(2) = .3$	$g_A(3) = .2$	

A Simple Discrete Example (Joint cont.)

The tables for the other two prior distributions are (of course) different from this first one. The joint distribution for the second prior, $g_B(\theta, x)$, is given in Table 4.

Table 4 Joint Distribution for X and θ Corresponding to the Second Prior for θ

	$\theta = 1$	$\theta = 2$	$\theta = 3$	
$x = 1$.12	.1	.03	$f_B(1) = .25$
$x = 2$.09	.1	.06	$f_B(2) = .25$
$x = 3$.06	.1	.09	$f_B(3) = .25$
$x = 4$.03	.1	.12	$f_B(4) = .25$
	$g_B(1) = .3$	$g_B(2) = .4$	$g_B(3) = .3$	

A Simple Discrete Example (Joint cont.)

The joint distribution for the third prior, $g_C(\theta, x)$, is given in Table 5.

Table 5 Joint Distribution for X and θ Corresponding to the Third Prior for θ

	$\theta = 1$	$\theta = 2$	$\theta = 3$	
$x = 1$.08	.075	.05	$f_C(1) = .205$
$x = 2$.06	.075	.1	$f_C(2) = .235$
$x = 3$.04	.075	.15	$f_C(3) = .265$
$x = 4$.02	.075	.2	$f_C(4) = .295$
	$g_C(1) = .2$	$g_C(2) = .3$	$g_C(3) = .5$	

A Simple Discrete Example (Posteriors)

The crux of the Bayes paradigm is that a *joint* probability distribution for X and θ , can be used not only to recover $f(x|\theta)$ and the marginal distribution of θ (the “likelihood” and the “prior distribution” that are multiplied together to get the joint distribution in the first place), but also to find conditional distributions for θ given possible values of X . In the context of Bayes analysis, these are called the *posterior* distributions of θ . For tables laid out as above, they are found by “dividing rows by row totals.” We might use the notation $g(\theta|x)$ for a posterior distribution and note that for a given x , values of this are proportional to joint probabilities, i.e.

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

i.e.

$$\text{posterior} \propto \text{likelihood} \cdot \text{prior}$$

Take, for example the situation of the first prior distribution.

A Simple Discrete Example (Posteriors cont.)

The four possible posterior distributions of θ (given the observed value of $X = x$) for the first prior distribution are as in Table 6.

Table 6 Posterior Distributions Corresponding to the First Prior for θ

	$\theta = 1$	$\theta = 2$	$\theta = 3$
$g_A(\theta 1)$	$.2/.295 =$.6780	$.075/.295 =$.2542	$.02/.295 =$.0678
$g_A(\theta 2)$	$.15/.265 =$.5660	$.075/.265 =$.2830	$.04/.265 =$.1509
$g_A(\theta 3)$.4255	.3191	.2553
$g_A(\theta 4)$.2439	.3659	.3902

A Simple Discrete Example (Posteriors cont.)

The set of posteriors for the second prior is in Table 7.

Table 7 Posterior Distributions Corresponding to the Second Prior for θ

	$\theta = 1$	$\theta = 2$	$\theta = 3$
$g_B(\theta 1)$.48	.40	.12
$g_B(\theta 2)$.36	.40	.24
$g_B(\theta 3)$.24	.40	.36
$g_B(\theta 4)$.12	.40	.48

A Simple Discrete Example (Posteriors cont.)

The set of posteriors for the third prior is in Table 8.

Table 8 Posterior Distributions Corresponding to the Third Prior for θ

	$\theta = 1$	$\theta = 2$	$\theta = 3$
$g_C(\theta 1)$.3902	.3659	.2439
$g_C(\theta 2)$.2553	.3191	.4255
$g_C(\theta 3)$.1509	.2830	.5660
$g_C(\theta 4)$.0678	.2542	.6780

A Simple Discrete Example (Posteriors cont.)

In all of these examples, as the observed value x increases, the posterior distribution shifts away from concentrating on “small” θ to concentrating on “large” θ . This should be consistent with intuition based on the form of the distributions of X in Table 1. But the strength of the “shift” depends upon the nature of the prior information employed. The posteriors in Table 6 are data-based modifications of the first prior of Table 2 (which favors small θ), the posteriors in Table 7 are data-based modifications of the second prior in Table 2 (which is relatively flat), and the posteriors in Table 8 are data-based modifications of the third prior in Table 2 (which favors large θ). The differences between Tables 6, 7, and 8 simply reflect the prior differences seen in Table 2.

A Simple Discrete Example (Inference)

The “Bayes paradigm” of inference is to base all formal inferences (plausibility statements) about θ on a posterior distribution of θ . For example, an analyst who adopts prior B and observes $X = 3$ may correctly say that there is (posterior) probability .36 that $\theta = 3$. Notice, that this is a different concept than the non-Bayesian concept of “confidence.”

A Simple Discrete Example (Prediction)

In many real contexts, one is doing inference based on data $X = x$ for the purpose of *predicting* the value of an as yet unobserved variable, X_{new} . If (given the value of θ) one is willing to model X and X_{new} as independent variables, one can extend the Bayes paradigm beyond inference for θ to the prediction problem. That is, conditioned on having observed $X = x$ one has a posterior distribution for both θ and X_{new} with a (joint) probability mass function that can be represented in a two way table, where each entry has the form

$$g(\theta, x_{\text{new}}|x) = f(x_{\text{new}}|\theta) g(\theta|x)$$

This can be added across values of θ to produce a *posterior predictive distribution* for X_{new} as

$$g(x_{\text{new}}|x) = \sum_{\theta} g(\theta, x_{\text{new}}|x) = \sum_{\theta} f(x_{\text{new}}|\theta) g(\theta|x)$$

A Simple Discrete Example (Prediction cont.)

That is, the posterior predictive distribution of X_{new} is what one gets upon weighting the three possible distributions for X_{new} in Table 1 according to the posterior probabilities in Tables 6, 7, or 8. For example, consider the case of the first prior of Table 2 and therefore the posterior distributions in Table 6. Considering first the possibility that $X = 1$, note that the conditional distribution for X_{new} given this outcome is

x_{new}	$g_A(x_{\text{new}} 1)$
1	$.4(.6780) + .25(.2542) + .1(.0678) = .34153$
2	$.3(.6780) + .25(.2542) + .2(.0678) = .28051$
3	$.2(.6780) + .25(.2542) + .3(.0678) = .21949$
4	$.1(.6780) + .25(.2542) + .4(.0678) = .15847$

The entire set of predictive distributions of X_{new} is then given in Table 9.

A Simple Discrete Example (Prediction cont.)

Table 9 Posterior Predictive Distributions for X_{new} Based on the First Prior for θ

x_{new}	$g_A(x_{\text{new}} 1)$
1	.34153
2	.28051
3	.21949
4	.15847

x_{new}	$g_A(x_{\text{new}} 2)$
1	.31224
2	.27073
3	.22922
4	.18771

x_{new}	$g_A(x_{\text{new}} 3)$
1	.27551
2	.25849
3	.24147
4	.22445

x_{new}	$g_A(x_{\text{new}} 4)$
1	.22806
2	.24269
3	.25732
4	.27195

A Simple Discrete Example (Prediction cont.)

The set of posterior predictive distributions for the second prior of Table 2 (and therefore the posterior distributions in Table 7) is given in Table 10.

Table 10 Posterior Predictive Distributions for X_{new} Based on the Second Prior for θ

x_{new}	$g_B(x_{\text{new}} 1)$
1	.30400
2	.26800
3	.23200
4	.19600

x_{new}	$g_B(x_{\text{new}} 2)$
1	.26800
2	.25600
3	.24400
4	.23200

x_{new}	$g_B(x_{\text{new}} 3)$
1	.23200
2	.24400
3	.25600
4	.26800

x_{new}	$g_B(x_{\text{new}} 4)$
1	.19600
2	.23200
3	.26800
4	.30400

A Simple Discrete Example (Prediction cont.)

The set of posterior predictive distributions for the third prior of Table 2 (and therefore the posterior distributions in Table 8) is given in Table 11.

Table 11 Posterior Predictive Distributions for X_{new} Based on the Third Prior for θ

x_{new}	$g_C(x_{\text{new}} 1)$
1	.27195
2	.25732
3	.24269
4	.22806

x_{new}	$g_C(x_{\text{new}} 2)$
1	.22445
2	.24147
3	.25849
4	.27551

x_{new}	$g_C(x_{\text{new}} 3)$
1	.18771
2	.22922
3	.27073
4	.31224

x_{new}	$g_C(x_{\text{new}} 4)$
1	.15847
2	.21949
3	.28051
4	.34153

A Simple Discrete Example (Prediction cont.)

For all 3 priors, the predictive distributions for X_{new} to some degree "follow x " in the sense that "small" x leads to a predictive distribution shifted toward "small" values of x_{new} , while "large" x shifts the predictive distribution toward "large" values of x_{new} . The prior buffers this effect, in that all of the posteriors for the first prior are in some sense shifted toward smaller values than for the other 2 priors (and all of the predictive distribution for the last prior are in some sense shifted toward larger values than for the other 2 priors). This is consistent with the story told in Table 2 about the form of the different prior assumptions.

Note too that in some sense the predictive distributions for X_{new} are qualitatively "flatter" than at least the first and last models for X and X_{new} represented in Table 1. This makes sense, as they are (posterior weighted) averages of the three distributions in that table.