Bayesian Statistical Reasoning: an Inferential, Predictive and Decision-Making Paradigm for the 21st Century

# **David Draper**

Department of Applied Mathematics and Statistics University of California, Santa Cruz, USA

draper@ams.ucsc.edu

www.ams.ucsc.edu/~draper

San Francisco Bay Area Chapter of the American Statistical Association Cal State East Bay, Hayward

19 November 2009

• **Statistics** is the study of **uncertainty**: how to **measure** it, and how to make **choices** in the face of it.

Since uncertainty is an **inescapable** part of the **human condition**, statistics has the potential to be **helpful** in almost every aspect of **daily life**, including **science** (the acquisition of knowledge for its own sake) and **decision-making** (how to use that knowledge to make a choice among the **available actions**).

When you notice you're **uncertain** about something — for example, the truth status of a **true-false proposition** such as "This patient is **HIV-positive**" or "Obama will win a **second term** as U.S. President in 2012" — it's natural

to want

(a) to quantify **how much uncertainty** you have and

(b) to figure out how to **reduce your uncertainty** if the answer to (a) is higher than the level necessary to achieve your goals.

**Probability** is the part of mathematics devoted to quantifying uncertainty, so it plays a **fundamental role** in statistics,

# Description, Inference, Prediction, ...

and so does **data-gathering**, because the best way to reduce your uncertainty is to get some relevant new information (data).

Statistical activities are of **four basic types**:

• **Description** of the important features of a data set, without an attempt to generalize outward from it (this activity is almost completely **non-probabilistic**, and I won't have much to say about it in this talk).

• Inference about the nature of the underlying **process** generating the data.

This is the statistical version of what the **18th-century** philosopher **Hume** referred to as the **problem of induction**; it includes as special cases (a) answering questions about **causality** and (b) generalizing outward from a **sample** of data values to a **population** (a broader universe of discourse).

• **Prediction** of future data on the basis of past data, including quantifying how much uncertainty you have about your predictions.

This is important in science, because good (bad) scientific theories make good (bad) predictions, and it's also important in

# **Decision-Making**; Frequentist, Bayesian Probability

• **Decision-making**: predicting the future under all the possible actions open to you, and choosing your favorite future on that basis.

The systematic study of **probability** can be traced back to an exchange of letters between **Pascal** and **Fermat** in the **1650s**, but the version of probability they developed turns out to be **too simplistic** to help in 21st-century problems of **realistic complexity**.

Instead, two other ways to give **meaning** to the **concept of probability** are in current use today:

• the **frequentist** (or relative-frequency) approach, in which you restrict attention to phenomena that are **inherently repeatable** under "identical" conditions and define P(A) to be the limiting relative frequency with which Awould occur in n repetitions, as  $n \to \infty$  (this approach was developed around **1870** by **Venn**, **Boole** and others and was refined in the **1930s** by **von Mises**); and

• the **Bayesian** approach, in which the argument B of the probability operator P(B|A) is a true-false proposition whose truth status is unknown to you

# Internal/External Information; Bayesian Computation

and P(B|A) represents the weight of evidence in favor of the truth of B, given the information in A (this approach was first developed by Bayes and Laplace in the 18th century and was refined by Keynes, de Finetti, Ramsay, Jeffreys, Turing, Good, Savage, Jaynes and others in the 20th century).
The Bayesian approach includes the frequentist paradigm as a special case, so you might think it would be the only version of probability used in statistical work today, but

• in quantifying your uncertainty about something unknown to you, the **Bayesian** paradigm requires you to bring **all relevant information** to bear on the calculation; this involves combining information both **internal** and **external** to the data set you've gathered, and (somewhat strangely) the **external**-information part of this approach was **controversial** in the 20th century, and

• Bayesian calculations require approximating high-dimensional integrals (whereas the frequentist approach mainly relies on maximization rather than integration), and this was a severe limitation to the Bayesian paradigm for a long time (from the 1750s to the 1980s).

Around **1990** two things happened roughly simultaneously that completely changed the **Bayesian computational** picture:

• Bayesian statisticians belatedly discovered that applied mathematicians (led by Metropolis and Ulam), working at the intersection between chemistry and physics in the 1940s, had used Markov chains to develop a clever algorithm, for approximating integrals arising in thermodynamics that are similar to the kinds of integrals that come up in Bayesian statistics, and

• desk-top computers finally became fast enough to implement the Metropolis algorithm in a feasibly short amount of time.

The 20th century was definitely a frequentist century, in large part because maximization was an excellent technology for that moment in history, from the 1920s (when the statistician and geneticist Fisher emphasized it) through the 1980s; but a consensus is now emerging around the idea that

 $\rightarrow$  In the 21st century it's important for statisticians to be fluent in both the frequentist and Bayesian ways of thinking.

In the **20th century** many people acted as if you had to **choose** one of these paradigms and **defend** it against attacks from people who favored the other one, but it turns out that **both approaches have strengths and weaknesses**, so that can't be the right way to **frame** the issue: it seems to me instead that my job as a statistician in this century is to develop a **fusion** of the two approaches that **emphasizes the strengths and de-emphasizes the weaknesses**.

#### My **personal fusion** involves

 reasoning in a Bayesian way when formulating my inferences, predictions and decisions, because the Bayesian paradigm is the most flexible
 approach so far developed for incorporating all relevant sources of uncertainty;

• reasoning in a **frequentist** way when paying attention to **how often I get the right answer**, which is an inherently **frequentist** activity that's **central** to good science and decision-making.

In this talk I'll (a) expand on the brief **historical notes** above and (b) give examples of **Bayesian** inference, prediction and decision-making in several case studies from **medicine** and **health policy**, illustrating the fusion just mentioned. • According to the useful history of mathematics web site www-history.mcs.st-and.ac.uk, mathematics began in Babylonia in approximately 2,000 BCE, with the development of a systematic way to record and manipulate numbers (both integers and fractions).

• Gambling, which you would think might prompt the creation of a mathematics based on what we now call **randomness**, is even **older**: dice-like objects made from animal bones have been traced back to at least 4,500 BCE.

• Thus we've been thinking mathematically as a species for about 4,000 years and gambling for far longer than that, and yet no one seems to have laid down the foundations of probability until around 350 years ago.

Some specialized problems in games of chance had been solved by Italian mathematicians going back to the 1400s, and Galileo Galilei (1564–1642) worked in a fragmentary way on probability concepts in the early 17th century, but the subject was not properly launched as a branch of mathematics until an exchange of letters between the French mathematicians Blaise Pascal (1623–1662) and Pierre de Fermat (1601–1665) in 1654.

• Pascal and Fermat invented what we now call the classical approach to probability: I enumerate the elemental outcomes (EOs) (the fundamental possibilities in the process under study) in a way that makes them equipossible (i.e., so that none would be favored over any other in hypothetical repetitions of the process) and compute the classical probability  $P_C(A)$  of an outcome A as the ratio of  $n_A$  = number of EOs favorable to A to n = total number of EOs.

This works for assigning probabilities to outcomes of idealized games of chance (dice, coins, roulette, cards) but fails in complicated problems like those people think about routinely today (e.g., what are the EOs in a regression setting with 100,000 observations and 1,000 predictor variables?).

• The Dutch scientist Christiaan Huygens (1629–1695) published the first book on probability in 1657.

• Another important early probability book was written by the Swiss mathematician Jacob **Bernoulli (1654–1705)** and published in 1713, after his death; in it Bernoulli stated and proved the first (weak) **law of large numbers**  $(\stackrel{P}{\rightarrow}$  of a sequence of random variables  $\bar{y}_n$  to a non-random limit  $\mu = E(y)$ ). • The **Pascal-Fermat classical approach** had no notion of **conditional probability**; this was remedied by **Thomas Bayes (1702–1761)**, who gave the first definition of

$$P(B|A) = \frac{P(B \text{ and } A)}{P(A)}, \text{ from which}$$
$$P(B \text{ and } A) = P(A) P(B|A) \tag{1}$$

for (true-false) propositions A and B, in a posthumous publication in 1764.
Bayes was interested in causal relationships: you see an effect in the world (e.g., people dying of a disease) and you wonder what was its cause (e.g., drinking the water? eating something? breathing the air? ...).

He had the bravery/imagination to consider this **probabilistically**, and he noticed that P(effect|cause) was a lot **easier** to think about than P(cause|effect), so he wondered how P(B|A) depended on P(A|B) (he wanted to **reverse the order of conditioning**).

#### **Bayes's Theorem for Propositions**

To find out he wrote down his definition in the other order:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}, \text{ from which}$$
$$P(A \text{ and } B) = P(B) P(A|B). \tag{2}$$

So **now** he has

 $P(B \text{ and } A) = P(A) P(B|A) \text{ and } P(A \text{ and } B) = P(B) P(A|B), \quad (3)$ 

and he can equate the two equations, since P(B and A) = P(A and B), and solve for what he wants to get

**Bayes's Theorem for propositions:**  $P(B|A) = \frac{P(B) P(A|B)}{P(A)}$ . (4)

The main application he had in mind was more ambitious: B stood for an **unknown rate** at which something happens (today we might use the symbol  $0 < \theta < 1$ ) and A stood for some **data** relevant to  $\theta$  (in today's notation his data set was  $y = (y_1, \ldots, y_n)$ , where each  $y_i$  was a 1/0 variable with success rate  $\theta$ ).

In words he thought of his result as having the following meaning:

$$P(\text{unknown}|\text{data}) = \frac{P(\text{unknown}) P(\text{data}|\text{unknown})}{P(\text{data})}.$$
 (5)

He conjectured (correctly) that his Theorem still applies when B is a real number  $(\theta)$  and A is a vector of real numbers (y); in contemporary notation

$$p(\theta|y) = \frac{p(\theta) \, p(y|\theta)}{p(y)},\tag{6}$$

where (a)  $p(\theta|y)$  and  $p(y|\theta)$  are **conditional probability densities** for  $\theta$  given yand y given  $\theta$  (respectively) and (b)  $p(\theta)$  and p(y) are **(unconditional) probability densities** for  $\theta$  and y (respectively).

This requires some **interpreting**: I want to use (6) **after** the data set y has arrived, to **quantify my uncertainty** about  $\theta$  in light of the new information, so I want to **condition on the data**, i.e., to treat the entire equation as a **function** of  $\theta$  for fixed y; this has two **implications**:

# Likelihood Function

(a) p(y) is just a **constant** — in fact, I can think of it as the **normalizing constant**, put into the equation to make the product  $p(\theta) p(y|\theta)$  **integrate to 1** (as all densities, e.g.,  $p(\theta|y)$ , must); and

(b)  $p(y|\theta)$  may look like the **sampling distribution** for y given  $\theta$ , but I have to think of it as a **function** of  $\theta$  for fixed y.

Much later, **Fisher (1922)** popularized this same idea and called it the **likelihood function** —

$$l(\theta|y) = c \, p(y|\theta),\tag{7}$$

where c is an arbitrary **positive constant** — but **Bayes** (1764) saw its importance first.

With this new notation and terminology **Bayes's Theorem** becomes

$$p(\theta|y) = c \, p(\theta) \, l(\theta|y). \tag{8}$$

 $l(\theta|y)$  represents the **information** about the unknown  $\theta$  **internal** to the data set y, but this is only **one ingredient** in the process of drawing together all of the **evidence** about  $\theta$ ;

## Synthesis of Knowledge

as Bayes (1764) understood, there will typically also be **information** about  $\theta$ **external** to y, and  $p(\theta)$  is where this other information comes into the **synthesis of knowledge**.

On the log scale, and ignoring irrelevant constants, Bayes's Theorem says

$$\ln p(\theta|y) = \ln p(\theta) + \ln l(\theta|y), \tag{9}$$

which, in words, could be **interpreted** as

$$\begin{pmatrix} \text{total information} \\ \text{about } \theta \end{pmatrix} = \begin{pmatrix} \text{information} \\ \text{external to } y \end{pmatrix} + \begin{pmatrix} \text{information} \\ \text{internal to } y \end{pmatrix}. \quad (10)$$

One way (but not the only way) you could think about the **information** about  $\theta$  external to y is to recall the **sequential** nature of learning: the temporal event of observing the data set y divides the time line into the period before y (a priori) and the period after y (a posteriori).

Centuries after Bayes, **Raiffa and Schlafer (1961)** used this to suggest a **different way** to express (9):

# Prior, Likelihood, Posterior

	$\ln p( heta y)$	=	$\ln p( heta)$	+	$\ln l( heta y)$	
(	posterior		( prior )		( likelihood )	
	information	=	information	+	information	•
	about $\theta$		$\left( about \theta \right)$		$\left( about \theta \right)$	
						(11)

With this in mind **Raiffa and Schlafer** called  $p(\theta|y)$  the **posterior distribution** and  $p(\theta)$  the **prior distribution** for  $\theta$ , respectively.

These are actually **not very good names**, because (as noted above)  $p(\theta|y)$  is meant to quantify all information about  $\theta$  **external** to y (whether that information arrives before or after y is irrelevant), but through widespread usage **we're stuck with them now**.

With this notation and terminology **Bayes's Theorem** says

$$p(\theta|y) = c \cdot p(\theta) \cdot l(\theta|y)$$
(posterior) = c \cdot (prior) \cdot (likelihood). (12)

## Laplace; Difficult Integrals Emerge

In theory, at least, the **posterior distribution**  $p(\theta|y)$  completely solves the **inference problem** about the unknown  $\theta$ , and **Bayes** already had figured this out in the **1760s**.

 History's second Bayesian, and a better Bayesian than Bayes (because he was a much better mathematician), was Pierre-Simon, Marquis de Laplace (1749–1827).

In the late 1700s Laplace independently re-discovered Bayes's Theorem and extended it to settings in which the unknown  $\theta$  was a vector of real numbers of length k; in this setting no changes are needed to the notation —

$$p(\theta|y) = c \, p(\theta) \, l(\theta|y) \tag{13}$$

— but now  $p(\theta|y), p(\theta)$  and  $l(\theta|y)$  are all probability **densities** on  $\mathbb{R}^k$  (if we want, we can choose c in  $l(\theta|y) = c p(y|\theta)$  to make  $l(\theta|y)$  a density).

Now, however, to evaluate c you need to compute a k-dimensional integral:

with 
$$\theta = (\theta_1, \dots, \theta_k), \quad c = \left(\int \dots \int p(\theta) \, l(\theta|y) \, d\theta_1 \dots d\theta_k\right)^{-1}.$$
 (14)

This is perhaps not so bad if k is 1 or 2, but already in Laplace's time he wanted to work on problems with  $k \ge 10$ ; moreover, even if you can compute c, for k > 2it's hard to **visualize** a k-dimensional posterior distribution  $p(\theta|y)$ , so you'll want to look at the k marginal distributions

$$p(\theta_j|y) = \int \cdots \int p(\theta|y) \, d\theta_1 \cdots d\theta_{j-1} d\theta_{j+1} \cdots d\theta_k, \tag{15}$$

and each of these involves a (k-1)-dimensional integral.

This — approximating high-dimensional integrals — is the Bayesian computation problem; remarkably, in 1774 Laplace developed a class of solutions, which we now refer to as Laplace approximations (based on Taylor series and the multivariate Gaussian distribution); even more remarkably, this method was forgotten after Laplace's death and was not independently re-discovered until the 1950s, where it re-emerged in the applied mathematics literature under the name saddlepoint approximations. All (or almost all) inferential work from 1764 to 1922 was Bayesian; for example, Gauss (1809), Galton (1888), Pearson (1892), and even Fisher (1915) reasoned completely in the Bayesian paradigm during this period.

• In 1866 Venn published *The Logic of Chance*, in which he introduced the frequentist approach to probability; this was part of a movement among scientists in Victorian England claiming that science should be objective (they believed that two scientists with the same data set should reach the same conclusions); the Bayesian imperative to combine information both internal (likelihood distribution) and external (prior distribution) to the data set bothered Venn, because if the two scientists had different external information they might reach different conclusions, and this went against his definition of objectivity (in computer science language, Venn called this a bug; Bayesians would call it a feature).

The **problem** with Venn's position, of course, is that **everything humans do is subjective** (based on **assumptions** and **judgments**); both **science** in general and **statistical inference** in particular are examples: (a) Good (bad) scientists exercise good (bad) judgment (that's how we know they're good (bad));

(b) All **probability** and **statistical modeling** in problems of **realistic complexity** involves **assumptions** and **judgments**.

Suppose, for example, that you and I and everybody else in the room are given a **big data set** (n = 10,000) where the outcome variable y is **{loan default or not}** and there are a lot (k = 500) of variables  $x_j$  (credit history) that may be useful in predicting loan status; we're all given a particular set of input values for the predictor variables and asked to work independently to predict P(default) for that individual.

There are so many judgment calls in building a model to do this (Which link function in the family of generalized linear models? How should the  $x_j$  enter the prediction process (linearly, quadratically, ...)? What subset of the  $x_j$  should be used? Which interactions among the  $x_j$  should be in the model? ...) that our estimates of P(default) could easily differ substantially, even though we all may be using the standard "objective" tools for model selection.

## Fisher's Version of the Likelihood Function

I believe the only reason Venn could have believed it was a good goal "that two scientists with the same data set should reach the same conclusions" was that he never did a complicated data analysis.

There's a **Bayesian** account of **objectivity**: to a Bayesian, saying that a probability is **objective** just means that **many reasonable people** would more or less **agree** on its value.

Since subjectivity is inevitable, the goal in statistical work should evidently be
(a) to make all of the assumptions and judgments in the analysis clear and
(b) to see how sensitive the conclusions are to reasonable perturbations in the assumptions and judgments.

• In 1922 Fisher recanted on his earlier Bayesian position — he had read Venn in the intervening years — and tried to create a non-Bayesian theory of inference without prior distributions, basing his theory on a frequentist interpretation of the likelihood function.

A simple example comparing **likelihood** and **Bayesian** modeling will help demonstrate how Fisher **did** and **did not succeed** in this attempt.

# **Example 1: Hospital Mortality**

**Example 1** (*hospital-specific prediction of mortality rates*): Suppose I'm interested in measuring the **quality of care** (e.g., Kahn et al., 1990) offered by one particular **hospital**.

I'm thinking of the **Dominican Hospital** (DH) in Santa Cruz, CA; if this were your problem you'd have a different hospital in mind.

As part of this I decide to examine the **medical records** of all patients treated at the DH in one particular time window, say **January 2006–December 2009**, for one particular **medical condition** for which there's a strong **process-outcome link**, say **acute myocardial infarction (AMI; heart attack)**.

(**Process** is what health care providers do on behalf of patients; **outcomes** are what happens as a result of that care.)

In the time window I'm interested in there will be about n = 400 **AMI patients** at the DH.

To keep things simple I'll ignore process for the moment and focus here on one particular outcome: **death status (mortality)** as of 30 days from hospital admission, coded 1 for dead and 0 for alive.

(In addition to process this will also depend on the **sickness at admission** of the AMI patients, but I'll ignore that initially too.)

From the vantage point of December 2005, say, **what may be said** about the roughly 400 1s and 0s I'll observe in 2006–09?

**Frequentist modeling.** By definition the frequentist approach is based on the idea of **hypothetical or actual repetitions** of the process being studied, under conditions that are as close to **independent identically distributed (IID)** sampling as possible.

When faced with a data set like the 400 1s and 0s  $(Y_1, \ldots, Y_n)$  here, the usual way to do this is to think of it **as a random sample**, or **like a random sample**, from some **population** that's of direct interest to me.

Then the **randomness** in my probability statements refers to the **process** of what I might get if I were to repeat the sampling over and over — the  $Y_i$  become **random variables** whose probability distribution is determined by this hypothetical repeated sampling.

#### **Independent Identically Distributed (IID) Sampling**

In the absence of any **predictor information** the off-the-shelf **frequentist model** for this situation is of course

$$Y_i \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta), \quad i = 1, \dots, n$$
 (16)

for some  $0 < \theta < 1$ , which plays the role of the **underlying mortality rate** in the **population** of patients to whom it's appropriate to **generalize outward** (what **IS** that population, by the way?): if  $\theta$  were unusually high, that would be **prima facie** evidence of a possible quality of care problem.

Since the  $Y_i$  are **independent**, the **joint** sampling distribution of all of them,  $P(Y_1 = y_1, \ldots, Y_n = y_n)$ , is the **product** of the separate, or **marginal**, sampling distributions  $P(Y_1 = y_1), \ldots, P(Y_n = y_n)$ :

$$P(Y_1 = y_1, \dots, Y_n = y_n) = P(Y_1 = y_1) \cdots P(Y_n = y_n)$$
$$= \prod_{i=1}^n P(Y_i = y_i).$$
(17)

But since the  $Y_i$  are also **identically distributed**, and each one is Bernoulli( $\theta$ ), i.e.,  $P(Y_i = y_i) = \theta^{y_i} (1 - \theta)^{1-y_i}$ , the joint sampling distribution can be written

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1 - y_i}.$$
 (18)

Let's use the symbol y to stand for the vector of **observed data values**  $(y_1, \ldots, y_n).$ 

Before any data have arrived, this joint sampling distribution is a function of y for fixed  $\theta$  — it tells me how the data would be likely to behave in the future if I were to take an IID sample from the Bernoulli( $\theta$ ) distribution.

In 1922 Fisher re-discovered the following idea (as noted earlier, Bayes and Laplace had it first): **after** the data have arrived it makes more sense to interpret (18) as a function of  $\theta$  for fixed y — Fisher called it the **likelihood function** for  $\theta$  in the Bernoulli( $\theta$ ) model:

$$l(\theta|y) = l(\theta|y_1, \dots, y_n) = \prod_{i=1}^n \theta^{y_i} (1-\theta)^{1-y_i}$$
(19)  
=  $P(Y_1 = y_1, \dots, Y_n = y_n)$  but interpreted  
as a function of  $\theta$  for fixed  $y_i$ .

#### Gaussian Likelihood With Large n

Fisher tried to create a theory of **inference** about  $\theta$  **based only on this function** — as noted above, this is an important ingredient, **but not the only important ingredient**, in inference from the Bayesian viewpoint. The Bernoulli( $\theta$ ) likelihood function can be **simplified** as follows:

$$l(\theta|y) = \theta^s (1-\theta)^{n-s}, \qquad (20)$$

where  $s = \sum_{i=1}^{n} y_i$  is the **number of 1s** in the sample and (n - s) is the **number of 0s**; what does this function **look like**, e.g., with n = 400 and s = 72 (this is similar to data you would get from the DH: a **30-day mortality rate** from AMI of **18%**)?



## Quadratic Log Likelihood With Large $\boldsymbol{n}$

Note that the likelihood function  $l(\theta|y) = \theta^s (1-\theta)^{n-s}$  in this problem **depends** on the data vector y only through  $s = \sum_{i=1}^n y_i$  — Fisher referred to any such data summary as a sufficient statistic (with respect to the assumed sampling model).

It's often at least as useful to look at the **logarithm** of the likelihood function as the likelihood function itself:



In this case, as is often true for large n, the log likelihood function looks **locally** quadratic around its maximum.

### Maximum Likelihood

Fisher had the further (frequentist) idea that the **maximum** of the likelihood function would be a good **estimate** of  $\theta$  (we'll look later at conditions under which this makes sense from the **Bayesian** viewpoint).

Since the logarithm function is monotone increasing, it's equivalent in maximizing the likelihood to **maximize the log likelihood**, and for a function as well behaved as this I can do that by setting its **first partial derivative** with respect to  $\theta$  to 0 and solving; here I get the familiar result

$$\hat{\theta}_{\rm MLE} = \frac{s}{n} = \bar{y}.\tag{21}$$

Fisher called the function of the data that **maximizes** the likelihood (or log likelihood) function the **maximum likelihood estimate** (MLE)  $\hat{\theta}_{MLE}$ .

Note also that if you maximize  $l(\theta|y)$  and I maximize  $c l(\theta|y)$  for any constant c > 0, we'll get the **same thing**, i.e., the likelihood function is only defined up to a **positive multiple**; Fisher's actual definition was

 $l(\theta|y) = c P(Y_1 = y_1, \dots, Y_n = y_n)$  for any (**normalizing constant**) c > 0.

**Frequentist inference:** (1) I think of my **data set** as **like** a **random sample** from some **population** (**challenge:** often **difficult** with **observational** data to **identify** what this population really is).

(2) I identify some **numerical summary**  $\theta$  of the population of interest (e.g., the **mean**), and I find a reasonable **estimate**  $\hat{\theta}$  of  $\theta$  based on the sample (**challenge:** how define **reasonable**?).

(3) I imagine repeating the random sampling, and I use the random behavior of  $\hat{\theta}$  across these hypothetical repetitions to make probability statements involving (but not about!)  $\theta$  (e.g., confidence intervals for  $\theta$  [e.g., "I'm 95% confident that  $\theta$  is between 0.14 and 0.22"] or hypothesis tests about  $\theta$  [e.g., the *P* value for testing  $H_0$ :  $\theta < 0.1$  against  $H_A$ :  $\theta \ge 0.1$  is near 0, so I reject  $H_0$ ).

I'm not allowed to make probability statements about  $\theta$  in the frequentist paradigm, because  $\theta$  is just a fixed unknown constant that's not changing across the hypothetical repetitions; thus  $P_F(0.14 < \theta < 0.22)$  is not meaningful, whereas  $P_B(0.14 < \theta < 0.22|y) \doteq 0.95$  makes perfect sense. From now on c in expressions like the likelihood function above will be a **generic** (and often **unspecified**) **positive constant**.

**Maximum likelihood** provides a basic principle for estimation of a (population) parameter  $\theta$  from the frequentist/likelihood point of view, but how should the **accuracy** of  $\hat{\theta}_{\text{MLE}}$  be assessed?

Evidently in the frequentist approach I want to compute the **variance** or **standard error** of  $\hat{\theta}_{\text{MLE}}$  in **repeated sampling**, or estimated versions of these quantities — I'll focus on the estimated variance  $\hat{V}(\hat{\theta}_{\text{MLE}})$ .

Fisher (1922) also proposed an **approximation** to  $\hat{V}(\hat{\theta}_{\text{MLE}})$  that works well for large *n* and makes **good intuitive sense**.

In the **AMI mortality** case study, where  $\hat{\theta}_{\text{MLE}} = \hat{\theta} = \frac{s}{n}$  (the **sample mean**), it's easy to show that

$$V(\hat{\theta}_{\rm MLE}) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad \hat{V}(\hat{\theta}_{\rm MLE}) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{22}$$

but Fisher wanted to derive results like this in a more **basic** and **general** way.

#### **Fisher Information**

In the language of this case study, Fisher noticed that if the sample size n increases while holding the MLE constant, the **second derivative of the log likelihood function at**  $\hat{\theta}_{\text{MLE}}$  (a negative number) **increases** in size.

This led him to define the **information** in the sample about  $\theta$  — in his honor it's now called the (observed) **Fisher information**:

$$\hat{I}\left(\hat{\theta}_{\rm MLE}\right) = \left[-\frac{\partial^2}{\partial\theta^2}\log l(\theta|y)\right]_{\theta=\hat{\theta}_{\rm MLE}}.$$
(23)

This quantity **increases** as n goes up, whereas my uncertainty about  $\theta$  based on the sample, as measured by  $\hat{V}(\hat{\theta}_{\text{MLE}})$ , should go **down** with n.

Fisher conjectured and proved that the information and the estimated variance of the MLE in repeated sampling have the following simple **inverse relationship** when n is large:

$$\hat{V}\left(\hat{\theta}_{\rm MLE}\right) \doteq \hat{I}^{-1}\left(\hat{\theta}_{\rm MLE}\right). \tag{24}$$

#### Likelihood-Based Large-Sample Confidence Intervals

In this case study the **Fisher information** and **repeated-sampling variance** come out

$$\hat{I}(\hat{\theta}_{\text{MLE}}) = \frac{n}{\hat{\theta}(1-\hat{\theta})} \quad \text{and} \quad \hat{V}(\hat{\theta}_{\text{MLE}}) = \frac{\hat{\theta}(1-\hat{\theta})}{n},$$
 (25)

which matches what I already know is **correct** in this case.

Fisher further proved that for large n (a) the MLE is approximately **unbiased**, meaning that in repeated sampling

$$E\left(\hat{\theta}_{\mathrm{MLE}}\right) \doteq \theta,$$
 (26)

and (b) the sampling distribution of the MLE is approximately **Gaussian** with mean  $\theta$  and estimated variance given by (24):

$$\hat{\theta}_{\text{MLE}} \stackrel{\cdot}{\sim} \text{Gaussian} \left[ \theta, \hat{I}^{-1} \left( \hat{\theta}_{\text{MLE}} \right) \right].$$
 (27)

Thus for large *n* an **approximate 95% confidence interval** for  $\theta$  is given by  $\hat{\theta}_{\text{MLE}} \pm 1.96 \sqrt{\hat{I}^{-1}(\hat{\theta}_{\text{MLE}})}.$ 

#### **Repeated-Sampling Asymptotic Optimality of MLE**

In the above expression for **Fisher information** in this problem,

$$\hat{I}\left(\hat{ heta}_{\mathrm{MLE}}
ight) = rac{n}{\hat{ heta}(1-\hat{ heta})},$$

as *n* increases,  $\hat{\theta}(1-\hat{\theta})$  will tend to the constant  $\theta(1-\theta)$  (this is well-defined because we've assumed that  $0 < \theta < 1$ , since  $\theta = 0$  and 1 are probabilistically uninteresting), which means that information about  $\theta$  on the basis of  $(y_1, \ldots, y_n)$ in the IID Bernoulli model **increases at a rate proportional to** *n* **as the sample size grows**.

This is generally true of the MLE (i.e., in regular parametric problems):

$$\hat{I}(\hat{\theta}_{\mathrm{MLE}}) = O(n) \quad \text{and} \quad \hat{V}(\hat{\theta}_{\mathrm{MLE}}) = O(n^{-1}), \qquad (28)$$

as  $n \to \infty$ , where the notation  $a_n = O(b_n)$  (as usual) means that the ratio  $\left|\frac{a_n}{b_n}\right|$  is bounded as n grows.

Thus uncertainty about  $\theta$  on the basis of the MLE goes down like  $\frac{c_{\text{MLE}}}{n}$  on the variance scale with more and more data (in fact Fisher showed that  $c_{\text{MLE}}$  achieves the lowest possible value: the MLE is efficient).

# **Bayesian Modeling**

As a Bayesian in this situation, my job is to quantify my uncertainty about the 400 binary **observables** I'll get to see starting in 2006, i.e., my initial modeling task is **predictive** rather than inferential.

There is no samples-and-populations story in this approach, but probability and random variables arise in a different way: quantifying my uncertainty (for the purpose of betting with someone about some aspect of the 1s and 0s, say) requires eliciting from myself a joint predictive distribution that accurately captures my judgments about what I'll see:  $P_{B:me}(Y_1 = y_1, \ldots, Y_n = y_n)$ .

Notice that in the frequentist approach the random variables describe the **process** of observing a repeatable event (the "random sampling" appealed to here), whereas in the Bayesian approach I use random variables to quantify **my uncertainty about observables I haven't seen yet**.

It turns out that the concept of probabilistic **accuracy** has two components: I want my uncertainty assessments to be both **internally** and **externally** consistent, which corresponds to the Bayesian and frequentist ideas of **coherence** and **calibration**, respectively.

# Exchangeability

Exchangeability as a Bayesian concept parallel to frequentist independence.

Eliciting a 400-dimensional distribution doesn't sound easy; major simplification is evidently needed.

In this case, and many others, this is provided by **exchangeability** considerations.

If (as in the frequentist approach) I have no relevant information that distinguishes one AMI patient from another, my uncertainty about the 400 1s and 0s is **symmetric**, in the sense that a random permutation of the **order** in which the 1s and 0s were labeled from 1 to 400 would leave my uncertainty about them unchanged.

de Finetti (1930, 1964) called random variables with this property **exchangeable**:

 $\{Y_i, i = 1, ..., n\}$  are **exchangeable** if the distributions of  $(Y_1, ..., Y_n)$  and  $(Y_{\pi(1)}, ..., Y_{\pi(n)})$  are the same for all permutations  $(\pi(1), ..., \pi(n))$ .

**NB** Exchangeability and IID are not the same: IID implies exchangeability, and exchangeable  $Y_i$  do have identical marginal distributions, but they're not independent (if I'm expecting **a priori** about 15% 1s, say (that's the 30-day death rate for AMI with average-quality care), the knowledge that in the first 50 outcomes at the DH 20 of them were deaths would certainly change my prediction of the 51st).

de Finetti also defined **partial** or **conditional exchangeability** (e.g., Draper et al., 1993): if, e.g., the gender X of the AMI patients were available, and if there were evidence from the medical literature that 1s tended to be noticeably more likely for men than women, then I would probably want to assume **conditional** exchangeability of the  $Y_i$  given X (meaning that the male and female 1s and 0s, viewed as separate collections of random variables, are each

#### unconditionally exchangeable).

This is related to Fisher's (1956) idea of recognizable subpopulations.

The judgment of exchangeability still seems to leave the joint distribution of the  $Y_i$  quite **imprecisely specified**.

After defining the concept of exchangeability, however, de Finetti went on to prove a **remarkable result**: if I'm willing to regard the  $\{Y_i, i = 1, ..., n\}$  as part (for instance, the beginning) of an **infinite** exchangeable sequence of 1s and 0s (meaning that every finite subsequence is exchangeable), then there's a simple way to characterize my joint predictive distribution, if it's to be **coherent** (e.g., de Finetti, 1975; Bernardo and Smith, 1994).

(Finite versions of the theorem have since been proven, which say that the longer the exchangeable sequence into which I'm willing to embed  $\{Y_i, i = 1, ..., n\}$ , the harder it becomes to achieve coherence with any probability specification that's far removed from the one below.)

**de Finetti's Representation Theorem.** If I'm willing to regard  $(Y_1, \ldots, Y_n)$  as the first *n* terms in an infinitely exchangeable binary sequence  $(Y_1, Y_2, \ldots)$ ; then, with  $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ ,

•  $\theta = \lim_{n \to \infty} \bar{Y}_n$  must exist, and the **marginal distribution** (given  $\theta$ ) for each of the  $Y_i$  must be  $P(Y_i = y_i | \theta) = \text{Bernoulli}(y_i | \theta) = \theta^{y_i} (1 - \theta)^{1 - y_i}$ ,
where P is my joint probability distribution on  $(Y_1, Y_2, ...)$ ;

•  $H(t) = \lim_{n \to \infty} P(\bar{Y}_n \leq t)$ , the **limiting cumulative distribution function** (CDF) of the  $\bar{Y}_n$  values, must also exist for all t and must be a valid CDF, and

•  $P(Y_1, \ldots, Y_n)$  can be expressed as

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i} (1-\theta)^{1-y_i} dH(\theta).$$
(29)

When (as will essentially always be the case in **realistic applications**) my joint distribution P is sufficiently **regular** that H possesses a **density** (with respect to Lebesgue measure),  $dH(\theta) = p(\theta) d\theta$ , (29) can be written in a **more accessible** 

way as

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \int_0^1 \theta^s (1 - \theta)^{n-s} p(\theta) \, d\theta,$$
(30)

where 
$$s = \sum_{i=1}^{n} y_i = n \bar{y}_n$$
.

#### The Law of Total Probability

$$P(Y_1 = y_1, \dots, Y_n = y_n) = p(y_1, \dots, y_n) = \int_0^1 \theta^s (1 - \theta)^{n-s} p(\theta) \, d\theta,$$

Now the Law of Total Probability says that, for all densities  $p(\theta)$ ,

$$p(y_1,\ldots,y_n) = \int_0^1 p(y|\theta) \, p(\theta) \, d\theta = \int_0^1 \theta^s (1-\theta)^{n-s} \, p(\theta) \, d\theta, \qquad (31)$$

This implies that in any **coherent** expression of uncertainty about **exchangeable** binary quantities  $Y_1, \ldots, Y_n$ ,

$$p(y_1, \dots, y_n | \theta) = \theta^s (1 - \theta)^{n-s}.$$
(32)

But (a) the left side of (32), interpreted as a function of  $\theta$  for fixed  $y = (y_1, \ldots, y_n)$ , is recognizable as the **likelihood function** for  $\theta$  given y, (b) the right side of (32) is recognizable as the likelihood function for  $\theta$  in **IID Bernoulli sampling**, and (c) (32) says that these must be the **same**.

Thus, to summarize de Finetti's Theorem **intuitively**, the assumption of exchangeability in my uncertainty about binary observables  $Y_1, \ldots, Y_n$  amounts to behaving **as if** 

# Mixture (Hierarchical) Modeling

- there is a quantity called  $\theta$ , interpretable as either the **long-run relative** frequency of 1s or the marginal probability that any of the  $Y_i$  is 1,
  - I need to treat  $\theta$  as a **random** quantity with density  $p(\theta)$ , and
    - conditional on this  $\theta$  the  $Y_i$  are IID Bernoulli( $\theta$ ).

In yet other words, for a Bayesian whose uncertainty about binary  $Y_i$  is exchangeable, the model may effectively be taken to have the simple **mixture** or **hierarchical** representation

$$\left\{\begin{array}{ccc} \theta & \sim & p(\theta) \\ (Y_i|\theta) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta), \ i = 1, \dots, n \end{array}\right\}.$$
 (33)

This is the **link** between frequentist and Bayesian modeling of binary outcomes: exchangeability implies that I should behave like a frequentist vis à vis the **likelihood function** (taking the  $Y_i$  to be IID Bernoulli( $\theta$ )), but a frequentist who treats  $\theta$  as a random variable with a **mixing distribution**  $p(\theta)$ .

To emphasize an important point mentioned above, to make sense of this in the Bayesian approach I have to treat  $\theta$  as a random variable, even though

# **Prior Specification**

logically it's a fixed unknown constant.

This is the main **conceptual** difference between the **Bayesian** and **frequentist** approaches: as a Bayesian I use the **machinery** of random variables to express my uncertainty about unknown quantities.

What's the **meaning** of the mixing distribution  $p(\theta)$ ?

 $p(\theta)$  doesn't involve  $y = (y_1, \ldots, y_n)$ , so it must represent my information about  $\theta$ external to the data set y; in other words, de Finetti's mixing distribution  $p(\theta)$ is Bayes's prior distribution.

# Example 1 (continued): Prior specification in the AMI mortality case study — let's say

(a) I know (from the literature) that the 30-day AMI mortality rate given average care and average sickness at admission in the U.S. is about 15%,

(b) I know little about care or patient sickness at the DH, but

(c) I'd be somewhat surprised if the "underlying rate" at the DH was much less than 5% or more than 30% (note the asymmetry).

#### The Beta Family of Densities on (0,1)

To quantify these judgments I seek a flexible family of densities on (0, 1), one of whose members has mean 0.15 and (say) 95% central interval (0.05, 0.30).

A convenient family for this purpose is the **beta** distributions,

$$Beta(\theta|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}, \qquad (34)$$

defined for  $(\alpha > 0, \beta > 0)$  and for  $0 < \theta < 1$ ; this family is **convenient** for two reasons: (1) It exhibits a wide variety of **distributional shapes**:



#### The Beta Family of Densities on (0,1)

As we saw above, the likelihood in this problem comes from the **Bernoulli** sampling distribution for the  $Y_i$ ,

$$p(y_1, \dots, y_n | \theta) = l(\theta | y) = \theta^s (1 - \theta)^{n-s}, \qquad (35)$$

where s is the sum of the  $y_i$ .

Now Bayes's Theorem says that to get the posterior distribution  $p(\theta|y)$  I **multiply** the prior  $p(\theta)$  and the likelihood — in this case  $\theta^s (1-\theta)^{n-s}$  — and **renormalize** so that the product integrates to 1.

Bayes himself noticed back in the 1750s that if the prior is taken to be of the form  $c \theta^u (1-\theta)^v$ , the product of the prior and the likelihood **will also be of this form**, which makes the **computations** more straightforward.

The beta family is said to be **conjugate** to the Bernoulli/binomial likelihood.

**Conjugacy** of a family of **prior** distributions to a given **likelihood** is a bit hard to define precisely, but the basic idea — given a particular likelihood function — is to try to find a family of prior distributions so that the **product** of members of this family with the likelihood function will also be in the family.

# The Beta Family (continued)

**Conjugate analysis** — finding conjugate priors for standard likelihoods and restricting attention to them on tractability grounds — is one of only two fairly general methods for getting closed-form answers in the Bayesian approach (the other is **asymptotic analysis**; see, e.g., Bernardo and Smith, 1994).

Suppose I restrict attention (for now) to members of the beta family in trying to specify a **prior distribution** for  $\theta$  in the AMI mortality example.

I want a member of this family which has mean 0.15 and 95% central interval (0.05, 0.30).

If  $\theta \sim \text{Beta}(\alpha, \beta)$ , it turns out that

$$E(\theta) = \frac{\alpha}{\alpha + \beta}$$
 and  $V(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$ . (36)

Setting  $\frac{\alpha}{\alpha+\beta} = 0.15$  and **solving** for  $\beta$  yields  $\beta = \frac{17}{3}\alpha$ ; then the equation

$$0.95 = \int_{0.05}^{0.30} \text{Beta}\left(\theta \left|\alpha, \frac{17}{3}\alpha\right.\right) \, d\theta \tag{37}$$

can readily be solved numerically for  $\alpha$  (e.g., in a symbolic computing

package such as Maple or a statistical computing package such as R) to yield  $(\alpha, \beta) = (4.5, 25.5)$ .



This prior distribution looks just like I want it to: it has a **long right-hand tail** and is **quite spread out**: the prior SD with this choice of  $(\alpha, \beta)$  is  $\sqrt{\frac{(4.5)(25.5)}{(4.5+25.5)^2(4.5+25.5+1)}} \doteq 0.064$ , i.e., my prior says that I think the underlying AMI mortality rate at the DH is around **15%**, give or take about **6 or 7%**.

#### **Hierarchical Model Expansion**

In the usual jargon  $\alpha$  and  $\beta$  are called **hyperparameters** since they're parameters of the prior distribution.

Written **hierarchically** the model I've arrived at is

 $(\alpha, \beta) = (4.5, 25.5) \quad (hyperparameters)$  $(\theta | \alpha, \beta) \sim Beta(\alpha, \beta) \quad (prior) \quad (38)$  $(Y_1, \dots, Y_n | \theta) \stackrel{\text{IID}}{\sim} Bernoulli(\theta) \quad (likelihood)$ 

(38) suggests what to do if I'm not sure about the specifications that led to  $(\alpha, \beta) = (4.5, 25.5)$ : hierarchically expand the model by placing a distribution on  $(\alpha, \beta)$  centered at (4.5, 25.5).

This is an important Bayesian modeling tool: if the model is inadequate in some way, **expand it hierarchically** in directions suggested by the nature of its inadequacy.

**Q:** Doesn't this set up the possibility of an **infinite regress**, i.e., how do I know **when to stop** adding layers to the hierarchy?

# **Conjugate Updating**

A: (1) In practice people stop when they run out of (time, money), after having made sure that the final model passes **diagnostic checks**; and comfort may be taken from the empirical fact that (2) there tends to be a kind of **diminishing** returns principle: the farther a given layer in the hierarchy is from the likelihood (data) layer, the less it tends to affect the answer.

The conjugacy of the prior leads to a **simple closed form** for the posterior here: with y as the vector of observed  $Y_i$ , i = 1, ..., n and s as the sum of the  $y_i$  (a **sufficient statistic** for  $\theta$ , as noted above, with the Bernoulli likelihood),

$$p(\theta|y,\alpha,\beta) = c \ l(\theta|y) \ p(\theta|\alpha,\beta)$$
  
=  $c \ \theta^s \ (1-\theta)^{n-s} \ \theta^{\alpha-1} (1-\theta)^{\beta-1}$  (39)  
=  $c \ \theta^{(s+\alpha)-1} (1-\theta)^{(n-s+\beta)-1},$ 

i.e., the **posterior** for  $\theta$  is  $Beta(\alpha + s, \beta + n - s)$ .

This gives the hyperparameters a useful interpretation in terms of **effective** information content of the prior: it's as if the data (Beta(s+1, n-s+1))were worth  $(s+1) + (n-s+1) \doteq n$  observations and the prior  $(\text{Beta}(\alpha, \beta))$  were worth  $(\alpha + \beta)$  observations.

## The Prior Data Set

This can be used to judge whether the prior is **more informative than** intended — here it's equivalent to (4.5 + 25.5) = 30 binary observables with a mean of 0.15.

In **Bayesian inference** the **prior information** can always be thought of as **equivalent** to a **prior data set**, in the sense that if

(a) I were to **merge** the **prior data set** with the **sample data set** and do a **likelihood analysis** on the **merged data**, and

(b) you were to do a **Bayesian analysis** with the **same prior information** and **likelihood**,

we would get the **same answers**.

Conjugate analysis has the advantage that the prior sample size can be explicitly worked out: here, for example, the **prior data set** in effect consists of  $\alpha = 4.5$  1s

and  $\beta = 25.5$  0s, with **prior sample size**  $n_0 = (\alpha + \beta) \doteq 30$ .

Even with **non-conjugate** Bayesian analyses, thinking of the **prior information** as equivalent to a **data set** is a **valuable heuristic**.

#### **Prior-To-Posterior Updating**

(39) can be **summarized** by saying

$$\left\{\begin{array}{c} \theta \sim \operatorname{Beta}(\alpha, \beta) \\ (Y_i|\theta) \stackrel{\text{IID}}{\sim} \operatorname{Bernoulli}(\theta), \\ i = 1, \dots, n \end{array}\right\} \to (\theta|y) \sim \operatorname{Beta}(\alpha + s, \beta + n - s), \quad (40)$$

where 
$$y = (y_1, ..., y_n)$$
 and  $s = \sum_{i=1}^n y_i$ .

Suppose the n = 400 DH patients include s = 72 deaths  $(\frac{s}{n} = 0.18)$ .



Then the **prior** is Beta(4.5, 25.5), the **likelihood** is Beta(73, 329), the **posterior** for  $\theta$  is Beta(76.5, 353.5), and the three densities plotted **on the same graph** are given above.

In this case the posterior and the likelihood nearly coincide, because the **data** information outweighs the prior information by  $\frac{400}{30}$  = more than 13 to 1. The mean of a Beta( $\alpha, \beta$ ) distribution is  $\frac{\alpha}{\alpha+\beta}$ ; with this in mind the posterior mean has an intuitive expression as a weighted average of the prior mean and data mean, with weights determined by the effective sample size of the prior,  $(\alpha + \beta)$ , and the data sample size n:

$$\frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right) \left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right) \left(\frac{s}{n}\right)$$
posterior
mean
178 = (.070) (.15) + (.93) (.18)

## **Comparison With Frequentist Modeling**

Another way to put this is that the data mean,  $\bar{y} = \frac{s}{n} = \frac{72}{400} = .18$ , has been **shrunk** toward the prior mean .15 by (in this case) a modest amount: the posterior mean is about .178, and the **shrinkage factor** is  $\frac{30}{30+400}$  = about .07.

**Comparison with frequentist modeling.** To analyze these data as a frequentist I would appeal to the **Central Limit Theorem:** n = 400 is big enough so that the repeated-sampling distribution of  $\bar{Y}$  is approximately  $N\left[\theta, \frac{\theta(1-\theta)}{n}\right]$ , so an approximate **95% confidence interval** for  $\theta$  would be centered at  $\hat{\theta} = \bar{y} = 0.18$ , with an estimated standard error of  $\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = 0.0192$ , and would run roughly from 0.142 to 0.218.

By contrast the posterior for  $\theta$  is also **approximately Gaussian** (see the graph on the next page), with a mean of 0.178 and an SD of

 $\sqrt{\frac{\alpha^*\beta^*}{(\alpha^*+\beta^*)^2(\alpha^*+\beta^*+1)}} = 0.0184$ , where  $\alpha^*$  and  $\beta^*$  are the parameters of the beta posterior distribution; a **95% central posterior interval** for  $\theta$  would then run from about 0.178 - (1.96)(0.0184) = 0.142 to 0.178 + (1.96)(0.0184) = 0.215.

The two approaches (frequentist based only on the sample, Bayesian based on the sample and the prior I'm using) give **almost the same** answers in this

# **Comparison With Frequentist Modeling (continued)**



case, a result that's typical of situations with fairly large n and relatively **diffuse** prior information.

Note, however, that the **interpretation** of the two analyses differs:

• In the frequentist approach  $\theta$  is fixed but unknown and  $\bar{Y}$  is random, with the analysis based on imagining what would happen if the hypothetical random sampling were repeated, and appealing to the fact that across these repetitions  $(\bar{Y} - \theta) \sim \text{Gaussian}(0, .019^2)$ ; whereas

# **Comparison With Frequentist Modeling (continued)**

In the Bayesian approach ȳ is fixed at its observed value and θ is treated as random, as a means of quantifying my posterior uncertainty about it: (θ − ȳ|ȳ) ∼ Gaussian(0, .018<sup>2</sup>).

This means among other things that, while it's **not legitimate** with the frequentist approach to say that  $P_F(.14 \le \theta \le .22) \doteq .95$ , which is what many users of confidence intervals would like them to mean, the corresponding statement  $P_B(.14 \le \theta \le .22|y, \text{diffuse prior information}) \doteq .95$  is a **natural consequence** of the Bayesian approach.

In the case of diffuse prior information and large n this justifies the fairly common informal practice of **computing inferential summaries in a frequentist way and then interpreting them in a Bayesian way**.

**Q:** When does **maximum likelihood** work well from a **Bayesian** viewpoint?

**A:** (i) When the **prior information** is **diffuse**, the **likelihood function** (interpreted as a density) and the **posterior distribution** will be **similar**; (ii) when the **sample size** n is **large**, both the **likelihood function** (interpreted as a density) and the **posterior distribution** will be close to (the same) **Gaussian**;

therefore when (i) and (ii) are true, **maximizing** over the likelihood function (frequentist) and **integrating** over it (Bayesian) will produce **similar answers**, and **differentiation** is easier than **integration**; so with a **large sample size** and **diffuse prior information** Fisher's technology provides a **convenient approximation** to the **Bayesian inferential answer**.

Some more history. • Fisher (1923) invents the analysis of variance for comparing the means of more than two samples, emphasizing P values from significance testing (in which you have a (sharp) null hypothesis (such as  $\theta = 0$ ) and no explicit alternative hypothesis).

• Ramsey (1926) invents Bayesian decision theory and shows that good (rational, coherent) decisions are found by maximizing expected utility.

• Neyman and Pearson (1928) — also working in the frequentist paradigm

— invent hypothesis testing, in which explicit null and alternative hypotheses are specified (such as  $H_0: \theta < 0.1$  versus  $H_A: \theta \ge 0.1$ ) and P values

play no part (instead you're supposed to define a rejection region in the sample space before the data are gathered and either reject the null or fail to reject it, depending on how the data come out).

# Exchangeability; Confidence Intervals; Metropolis Algorithm

- de Finetti (1930, 1938) defines exchangeability and demonstrates its central role in Bayesian modeling.
  - Neyman (1937) invents confidence intervals.
- Metropolis and Ulam (1949) define the Monte Carlo method and point out that anything you want to know about a probability distribution, no matter how complicated or high-dimensional, can be learned to arbitrary accuracy by sampling from it.

• Wald (1950) tries to create a frequentist decision theory to compete with Ramsey's Bayesian approach and finds, to his dismay, that all good decision rules are Bayes rules.

• Metropolis et al. (1953) publish the Metropolis algorithm, which solves the Bayesian computational problem (of approximating high-dimensional integrals); no one notices this fact.

• Savage (1954) publishes *The Foundations of Statistics*, in which he begins by trying to put **frequentist inference** on a **sound theoretical footing** and ends by concluding that this is **not possible**; the experience of writing the book

turns Savage into a **Bayesian**.

 Lindley (1965) publishes Introduction to Probability and Statistics From a Bayesian Viewpoint, in which he shows that (a) some popular frequentist inferential tools (e.g., confidence intervals) sometimes have approximate Bayesian interpretations but (b) others (e.g., P values) do not.

• Hastings (1970) generalizes the Metropolis algorithm and publishes the result in *Biometrika*; Bayesians still take no notice.

• Geman and Geman (1984) independently re-invent a special case of the Metropolis-Hastings algorithm, name it Gibbs sampling, and apply it to Bayesian image restoration; Bayesians not working in image restoration still

#### are **unaware**.

• Gelfand and Smith (1990) finally publicize Gibbs sampling in a mainstream statistics journal as a possible solution to the Bayesian computational problem, and desktop computers finally become fast enough to permit the algorithm to produce useful answers in small and medium-sized problems in under 12 hours of clock time; Bayesian applied statistics is now finally fully operational.

#### **Summary of the Bayesian Statistical Paradigm**

Three basic ingredients of the Bayesian statistical paradigm:

- $\theta$ , something of **interest** which is **unknown** (or only partially known) to me (e.g.,  $\theta_{RR}$ , the relative risk of getting a disease under one treatment condition versus another).
- Often  $\theta$  is a **parameter vector** (of finite length k, say) or a **matrix**, but it can literally be **almost anything**, e.g., a **function** (e.g., a **cumulative distribution function** (CDF) or **density**, a **regression surface**, ...), a **phylogenetic tree**, an **image** of the (true) surface of Mars, ... .
- y, an information source which is relevant to decreasing my uncertainty about  $\theta$ .

Often y is a vector of real numbers (of length n, say), but it can also literally be almost anything, e.g., a time series, a movie, the text in a book, ....

• A desire to learn about  $\theta$  from y in a way that is both coherent (internally consistent, i.e., free of internal logical contradictions) and well-calibrated (externally consistent, e.g., capable of making accurate predictions of future data  $y^*$ ).

# All Uncertainty Quantified With Probability Distributions

It turns out (e.g., de Finetti 1990, Jaynes 2003) that I'm compelled in this situation to reason within the standard rules of probability as the basis of my inferences about  $\theta$ , predictions of future data  $y^*$ , and decisions in the face of uncertainty, and to quantify my uncertainty about any unknown quantities through conditional probability distributions, as follows:

$$p(\theta|y, \mathcal{B}) = c \, p(\theta|\mathcal{B}) \, l(\theta|y, \mathcal{B})$$

$$p(y^*|y, \mathcal{B}) = \int p(y^*|\theta, \mathcal{B}) \, p(\theta|y, \mathcal{B}) \, d\theta$$

$$a^* = \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y, \mathcal{B})} \left[ U(a, \theta) \right]$$
(41)

•  $\mathcal{B}$  stands for my **background** (often not fully stated) **assumptions** and **judgments** about how the world works, as these assumptions relate to **learning** about  $\theta$  from y.

 $\mathcal{B} \text{ is often omitted from the basic equations (sometimes with unfortunate consequences), yielding the simpler-looking forms
<math display="block">p(\theta|y) = c \, p(\theta) \, l(\theta|y) \qquad p(y^*|y) = \int p(y^*|\theta) \, p(\theta|y) \, d\theta$   $a^* = \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y)} \left[ U(a, \theta) \right]$ (42)

#### **Prior and Posterior Distributions**

 $p(\theta|y, \mathcal{B}) = c \, p(\theta|\mathcal{B}) \, l(\theta|y, \mathcal{B}) \qquad p(y^*|y, \mathcal{B}) = \int p(y^*|\theta, \mathcal{B}) \, p(\theta|y, \mathcal{B}) \, d\theta$  $a^* = \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y, \mathcal{B})} \left[ U(a, \theta) \right]$ 

•  $p(\theta|\mathcal{B})$  is my (so-called) prior information about  $\theta$  given  $\mathcal{B}$ , in the form of a probability density function (PDF) or probability mass function (PMF) if  $\theta$  lives continuously or discretely on  $\mathbb{R}^k$  (let's just agree to call this my prior distribution), and  $p(\theta|y, \mathcal{B})$  is my (so-called) posterior distribution about  $\theta$  given y and  $\mathcal{B}$ , which summarizes my current total information about  $\theta$  and solves the inference problem.

These are actually **not very good names** for  $p(\theta|B)$  and  $p(\theta|y,B)$ , because (e.g.)  $p(\theta|B)$  really stands for **all (relevant) information** about  $\theta$  (given B) **external to** y, whether that information was obtained **before** (or **after**) yarrives, but (a) they do emphasize the **sequential nature of learning** and (b) through long usage **we're stuck with them**.

 c (here and throughout) is a generic positive normalizing constant, inserted into the first equation above to make the left-hand side integrate (or sum) to 1 (as any coherent distribution must).

# Sampling Distributions, Likelihood Functions and Utility

 $\begin{aligned} p(\theta|y,\mathcal{B}) &= c \, p(\theta|\mathcal{B}) \, l(\theta|y,\mathcal{B}) \qquad p(y^*|y,\mathcal{B}) = \int p(y^*|\theta,\mathcal{B}) \, p(\theta|y,\mathcal{B}) \, d\theta \\ a^* &= \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y,\mathcal{B})} \left[ U(a,\theta) \right] \end{aligned}$ 

•  $p(y^*|\theta, \mathcal{B})$  is my sampling distribution for future data values  $y^*$  given  $\theta$ and  $\mathcal{B}$  (and presumably I would use the same sampling distribution  $p(y|\theta, \mathcal{B})$ for (past) data values y, thinking before the data arrives about what values of y I might see).

This assumes that I'm willing to **regard** my data as **like random draws from a population of possible data values** (an **heroic assumption** in some cases, e.g., with **observational** rather than **randomized** data).

*l*(*θ*|*y*, *B*) is my likelihood function for *θ* given *y* and *B*, which is defined to be any positive constant multiple of the sampling distribution *p*(*y*|*θ*, *B*) but re-interpreted as a function of *θ* for fixed *y*:

$$l(\theta|y,\mathcal{B}) = c \, p(y|\theta,\mathcal{B}). \tag{43}$$

•  $\mathcal{A}$  is my set of possible **actions**,  $U(a, \theta)$  is the numerical value (**utility**) I attach to taking **action** a if the **unknown** is really  $\theta$ , and the third equation says I should **find** the action  $a^*$  that **maximizes expected utility** (MEU).

## **Predictive Distributions and MCMC**

 $p(\theta|y, \mathcal{B}) = c \, p(\theta|\mathcal{B}) \, l(\theta|y, \mathcal{B}) \qquad p(y^*|y, \mathcal{B}) = \int p(y^*|\theta, \mathcal{B}) \, p(\theta|y, \mathcal{B}) \, d\theta$  $a^* = \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y, \mathcal{B})} \left[ U(a, \theta) \right]$ 

And p(y\*|y, B), my (posterior) predictive distribution for future data y\* given (past) data y and B, must be a weighted average of my sampling distribution p(y\*|θ, B) weighted by my current best information p(θ|y, B) about θ given y and B.

That's the paradigm, and it's been highly successful in the past (say) 30 years — in fields as far-ranging as bioinformatics, econometrics, environmetrics, and medicine — at quantifying uncertainty in a coherent and well-calibrated way and helping people find satisfying answers to hard scientific questions.

Evaluating (potentially high-dimensional) integrals (like the one in the second equation above, and many others that arise in the Bayesian approach) is a technical challenge, often addressed these days with sampling-based
Markov chain Monte Carlo (MCMC) methods (e.g., Gilks, Richardson and Spiegelhalter 1996).

Quality of hospital care is often studied with cluster samples: I take a random sample of J hospitals (indexed by j) and a random sample of N total patients (indexed by i) nested in the chosen hospitals, and I measure quality of care for the chosen patients and various hospital- and patient-level predictors.

With  $y_{ij}$  as the **quality of care score** for patient *i* in hospital *j*, a first step would often be to fit a **variance-components model** with **random effects** at both the **hospital** and **patient** levels:

$$y_{ij} = \beta_0 + u_j + e_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, J;$$
  
$$\sum_{j=1}^J n_j = N, \quad (u_j | \sigma_u^2) \stackrel{\text{IID}}{\sim} N(0, \sigma_u^2), \quad (e_{ij} | \sigma_e^2) \stackrel{\text{IID}}{\sim} N(0, \sigma_e^2).$$
(44)

Browne and Draper (2006) used a simulation study to show that, with a variety of maximum-likelihood-based methods for creating confidence intervals for  $\sigma_u^2$ , the actual coverage of nominal 95% intervals ranged from 72% to 94% across realistic sample sizes and true parameter values, versus 89–94% for Bayesian methods.

## **Poorly-Calibrated Frequentist Inference (continued)**

In a re-analysis of a Guatemalan National Survey of Maternal and Child Health, with three-level data (births within mothers within communities), working with the random-effects logistic regression model

$$(y_{ijk} | p_{ijk}) \stackrel{\text{indep}}{\sim} \text{Bernoulli}(p_{ijk}) \quad \text{with}$$

$$\log it(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + v_k,$$

$$(45)$$

where  $y_{ijk}$  is a binary indicator of modern prenatal care or not and where  $u_{jk} \sim N(0, \sigma_u^2)$  and  $v_k \sim N(0, \sigma_v^2)$  were random effects at the mother and community levels (respectively), Browne and Draper (2006) showed that things can be even worse for likelihood-based methods, with actual coverages (at nominal 95%) as low as 0–2% for intervals for  $\sigma_u^2$  and  $\sigma_v^2$ , whereas Bayesian methods again produce actual coverages from 89–96%.

The technical problem is that the marginal likelihood functions for random-effects variances are often heavily skewed, with maxima at or near 0 even when the true variance is positive; Bayesian methods, which integrate over the likelihood function rather than maximizing it, can have (much) better small-sample calibration performance.

# HIV–1 Vaccine Efficacy

Two concluding points for this talk: (1) Inference and decision-making are not the same thing. (2) People sometimes use inferential tools to make an implied decision when decision-making methods lead to a better choice.

**Example 2:** A randomized controlled trial of an rgp120 vaccine against HIV (rgp120 HIV Vaccine Study Group (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV–1 infection. *Journal of Infectious Diseases*, **191**, 654–663).

5403 healthy HIV-negative volunteers at high risk of getting HIV were randomized, 3598 to the vaccine and 1805 to placebo (in both cases, 7 injections over 30 months), and followed for 36 months; the main outcome was presence or absence of HIV infection at the end of the trial, with Vaccine Efficacy (VE) defined as

 $VE = 100(1 - \text{ relative risk of infection}) = 100 \left[1 - \frac{P(\text{infection}|\text{vaccine})}{P(\text{infection}|\text{placebo})}\right].$ 

Secondary frequentist analyses examined differences in VE by gender, ethnicity, age, and education and behavioral risk score at baseline.

A reminder of how **frequentist hypothesis tests** work: e.g., to test  $H_0: \theta_{RR} < 1$  against  $H_A: \theta_{RR} \ge 1$  based on a **sample** of size *n*, the **optimal test** is of the form

**reject**  $H_0$  if  $\hat{\theta}_{RR} \ge c$ , where c is chosen to make

 $P_F(\text{type I error}) = P_F(\text{reject } H_0 \text{ when } H_0 \text{ is true}) \leq \alpha,$ 

in which  $\alpha$  is typically some **conventional value** like **0.05**; or **equivalently** you can **reject**  $H_0$  if

P value =  $P_F$ (getting data as extreme as, or more extreme than, what you got, if  $H_0$  is true)  $\leq \alpha$ .

If you have **control** over the **sample** size (e.g., at the time the **experiment** is **designed**), n is typically chosen so that

 $P_F(\text{type II error}) = P_F(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}) \leq \beta$ 

(subject to the constraint  $P_F(\text{type I error}) \leq \alpha$ ), in which  $\beta$  is typically some conventional value like 0.2  $(1 - \beta = \text{power} = 0.8)$ ; if you don't have control over n, typically only type I error is paid attention to.

# Vaccine Efficacy

	Rate	(%) of				
	HIV–1 Infection		VE	P Value		
Group	Vaccine	Placebo	(95%  CI)	Unadj	Adj	D-M
All	241/3598	127/1805	6 (-17)	50	<u> </u>	
Volunteers	(6.7)	(7.0)	to $24)$	.09	∕.0	
Black	6/233	9/116	67~(6	028	.24	
(Non-Hisp)	(2.6)	(7.8)	to 88)	.028		
Black	1/112	4/57	87 (19)	033		
Women	(0.9)	(7.0)	to 98)	.000		
Nonwhite	30/604	29/310	47 (12)	012 13		
Nonwinte	(5.0)	(9.4)	to 68)	.012	.10	
Nonwhite	27/461	25/236	43 (3	.036		
Men	(6.1)	(10.6)	to $67)$			

The trial found a small decline in infection overall (6.7% vaccine, 7.0% placebo) that was neither practically nor statistically significant; large preventive effects of the vaccine were found for some subgroups (e.g., nonwhites), but statistical significance vanished after adjustment for multiple comparisons.

#### **Frequentist Multiple Comparisons Adjustment**

	Rate					
	HIV–1 Infection		VE	P Value		
Group	Vaccine	Placebo	(95%  CI)	Unadj	Adj	D-M
Normhite	30/604	29/310	47 (12	019	12	
nonwinte	(5.0)	(9.4)	to $68)$	.012	.10	

Note that the *P* value for the nonwhite subgroup was 0.012 before, but 0.13 after, (frequentist) multiple comparisons adjustment.

However, frequentist multiple comparisons methods are an inferential approach to what should really be a decision problem (Should this vaccine be given to nonwhite people at high risk of getting HIV? Should another trial focusing on nonwhites be run?), and when multiple comparison methods are viewed as "solutions" to a Bayesian decision problem they do not have a sensible implied utility structure: they're terrified of announcing that an effect is real when it's not (a type I error), and have no built-in penalty for failing to announce an effect is real when it is (a type II error).

## **Decision-Making**

In the **frequentist** approach, **type II errors** are supposed to be **taken care of** by having done a **power calculation** at the time the **experiment** was **designed**, but this **begs the question** of **what decision should be taken**, **now that this study has been run**, about whether to **run a new trial** and/or **give the vaccine to nonwhite people now**.

When the problem is reformulated as a decision that properly weighs all of the real-world costs and benefits, the result (interpreted in frequentist language) would be a third P value column in the table on page 4 (a column called "Implied P from a decision-making perspective", or D-M for short) that would look a lot more like the first (unadjusted) P value column than the second (multiple-comparisons adjusted) column, leading to the decision that a new trial for nonwhites for this vaccine is a good clinical and health policy choice.

The point is that when the **problem** is really to **make a decision**, **decision-theoretic methods** typically lead to **better choices** than **inferential methods** that were **not intended to be used** in this way.

# **Decision-Theoretic Re-Analysis**

	Rate $(\%)$ of					
	HIV–1 Infection		VE	P Value		e
Group	Vaccine	Placebo	(95% CI)	Unadj	Adj	D-M
All	241/3598	127/1805	6 (-17)	.59	> .5	А
Volunteers	(6.7)	(7.0)	to $24)$			Lot
Black	6/233	9/116	67~(6	.028	8.24	More
(Non-Hisp)	(2.6)	(7.8)	to 88)			Like
Black	1/112	4/57	87 (19	.033	Т	The
Women	(0.9)	(7.0)	to $98)$			THE
Neurophite	30/604	29/310	47 (12)	.012	.13	Unodi
Nonwhite	(5.0)	(9.4)	to 68)			Unauj
Nonwhite	27/461	25/236	43 (3	.036		Col
Men	(6.1)	(10.6)	to $67)$			COI

When both type I and type II losses are properly traded off against each other (and gains are correctly factored in as well), the right choice is (at a minimum) to run a new trial in which Nonwhites (principally Blacks and Asians, both men and women) are the primary study group.

#### Details

This can be seen in an even simpler setting: consider a randomized controlled Phase 3 clinical trial with no subgroup analysis, and define  $\Delta$ to be the population mean health improvement from the treatment T as compared with the control condition C.

There will typically be some point c along the number line (a kind of practical significance threshold), which may not be 0, such that if  $\Delta \geq c$  the treatment should be implemented (note that this is really a decision problem, with action space  $a_1 = \{\text{implement } T\}$  and  $a_2 = \{\text{don't}\}$ ).

The frequentist hypothesis-testing inferential approach to this problem would test  $H_0: \Delta < c$  against  $H_A: \Delta \geq c$ , with (reject  $H_0$ ) corresponding to action  $a_1$ .

In the **frequentist inferential approach**  $H_0$  would be rejected if  $\hat{\Delta} \geq \Delta^*$ , where  $\hat{\Delta}$  is a **good estimator** of  $\Delta$  based on **clinical trial data** D with **sample size** n and  $\Delta^*$  is chosen so that the corresponding P value is no greater than  $\alpha$ , the **type I error probability** (the chance of **rejecting**  $H_0$  when  $H_0$  is **true**).

# **Details** (continued)

As noted above,  $\alpha$  is usually chosen to be a **conventional value** such as **0.05**, in conjunction with choosing *n* large enough (if you can do this at **design time**) so that the **type II error probability**  $\beta$  is no more than **another conventional value** such as **0.2** (the **real-world consequences** of **type I** and **type II** errors are rarely contemplated in choosing  $\alpha$  and  $\beta$ , and in practice you won't necessarily have a large enough *n* for, e.g., subgroup analyses to correctly control the **type II error probability**).

The Bayesian decision-theoretic approach to this decision problem requires me to specify a utility function that addresses these real-world consequences (and others as well); a realistic utility structure here would depend continuously on  $\Delta$ , but I can look at an oversimplified utility structure that permits comparison with hypothesis-testing: for  $u_{ij} \geq 0$ ,

	Truth			
Action	$\Delta \ge c$	$\Delta < c$		
$a_1$	$u_{11}$	$-u_{12}$		
$a_2$	$-u_{21}$	$u_{22}$		

# **Details** (continued)



The utilities may be considered from the point of view of several different actors in the drama; in the context of the HIV vaccine study, for instance, considering the situation from the viewpoint of a non-HIV+ person at high risk of becoming HIV+,

- $u_{11}$  is the **gain** from **using** a vaccine that is **thought** to be **effective** and **really is effective**;
- $-u_{12}$  is the loss from using a vaccine that is **thought** to be **effective** and really is not effective;
  - $-u_{21}$  is the loss from not using a vaccine that is **thought** to be not effective but really is effective; and
- $u_{22}$  is the gain from not using a vaccine that is thought to be not effective and really is not effective (i.e.,  $u_{22} = 0$ ).

## **Details** (continued)

Note that the **frequentist inferential approach** at **analysis time** only requires you to think about something ( $\alpha$ ) corresponding to **one** of these **four ingredients** ( $-u_{12}$ ), and even then  $\alpha$  is on the **wrong (probability) scale** (the  $u_{ij}$  will be on a **real-world-relevant scale** such as **quality-adjusted life years** (QALYs)).

The optimal Bayesian decision turns out to be

choose  $a_1$  (implement T)  $\leftrightarrow P(\Delta \ge c|D) \ge \frac{u_{12} + u_{22}}{u_{11} + u_{12} + u_{21} + u_{22}} = u^*$ .

The **frequentist inferential approach** is **equivalent** to this **only if** 

$$\alpha = 1 - u^* = \frac{u_{11} + u_{21}}{u_{11} + u_{12} + u_{21} + u_{22}}.$$

In the context of the **HIV vaccine**, with realistic values of the  $u_{ij}$  that appropriately weigh both the loss from taking the vaccine when it doesn't work and failing to take the vaccine when it does work, the analogous frequentist inferential "action" would be to reject  $H_0$  for Pvalues that are much larger than the usual threshold (e.g., **0.3** instead of **0.05**).