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

National University of Ireland, Galway

1 June 2010

Statistics; Probability

- $\boxed{\textbf{Statistics}}$ 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 A would occur in n repetitions, as $n \rightarrow \infty$ (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).

Metropolis Algorithm; Bayesian + Frequentist

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

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

Bayesian-Frequentist Fusion

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.

History of Probability and Statistics

- 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.

Classical Approach; Law of Large Numbers

- **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 ($\xrightarrow{P}$ of a sequence of random variables $\bar{y}_n$ to a non-random limit $\mu = E(y)$).

Conditional Probability

- 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

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

$$P(B \text{ and } A) = P(A)\, P(B|A) \qquad (1)$$

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 $P(\text{effect}|\text{cause})$ was a lot **easier** to think about than $P(\text{cause}|\text{effect})$, 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

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

and he can **equate the two equations**, since $P(B \text{ and } A) = P(A \text{ and } B)$, and **solve** for what he wants to get

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

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 θ (in today's notation his data set was $y = (y_1, \dots, y_n)$, where each y_i was a **1/0 variable** with success rate θ).

Bayes's Theorem for Real Numbers

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

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

He **conjectured** (correctly) that his Theorem still applies when B is a **real number** (θ) 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 θ given y and y given θ (respectively) and (b) $p(\theta)$ and $p(y)$ are **(unconditional) probability densities** for θ and y (respectively).

This requires some **interpreting**: I want to use (6) **after** the data set y has arrived, to **quantify my uncertainty** about θ 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 θ 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 θ, but I have to think of it as a **function** of θ 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 θ **internal** to the data set y, but this is only **one ingredient** in the process of drawing together all of the **evidence** about θ;

Synthesis of Knowledge

as Bayes (1764) understood, there will typically also be **information** about θ **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

$$\left(\begin{array}{c} \text{total information} \\ \text{about } \theta \end{array} \right) = \left(\begin{array}{c} \text{information} \\ \text{external to } y \end{array} \right) + \left(\begin{array}{c} \text{information} \\ \text{internal to } y \end{array} \right) . \tag{10}$$

One way (but not the only way) you could think about the **information** about θ 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, researchers in the **1950s** used this to suggest a **different way** to express (9):

Prior, Likelihood, Posterior

$$\begin{array}{ccccc} \ln p(\theta|y) & = & \ln p(\theta) & + & \ln l(\theta|y) \\ \left(\begin{array}{c} \text{posterior} \\ \text{information} \\ \text{about } \theta \end{array} \right) & = & \left(\begin{array}{c} \text{prior} \\ \text{information} \\ \text{about } \theta \end{array} \right) & + & \left(\begin{array}{c} \text{likelihood} \\ \text{information} \\ \text{about } \theta \end{array} \right) . \end{array} \tag{11}$$

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

These are actually **not very good names**, because (as noted above) $p(\theta|y)$ is meant to quantify all information about θ **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

$$\begin{array}{ccccccc} p(\theta|y) & = & c & \cdot & p(\theta) & \cdot & l(\theta|y) \\ (\text{ posterior }) & = & c & \cdot & (\text{ prior }) & \cdot & (\text{ likelihood }) . \end{array} \tag{12}$$

Prior and Likelihood Specification; Parametric Modeling

This creates a specification problem: how do you quantify "information about θ **internal** to y" in the **likelihood distribution** $l(\theta|y)$ and "information about θ **external** to y" in the **prior distribution** $p(\theta)$?

I'll give an example later of **prior specification**; what about specifying

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

From a Bayesian perspective $p(y|\theta)$ is the **predictive distribution** for how the data will come out **before any data have arrived**; how do you specify this?

Typical solution from **1764** through **1937**: try to find a **standard parametric family** of probability distributions (indexed by $\gamma = (\theta, \eta)$) that captures **what you expect to see** in the data (based on **previous experience** with similar problems); for example, with **binary outcomes** you would first try the **Bernoulli**(θ) distribution, with **count data** you would first think of the **Poisson**(θ) distribution, and with **continuous outcomes** you might well start with the **Normal**(θ, σ^2) distribution.

This — **parametric statistical modeling** — was the standard approach for centuries, but there's a **problem** with it:

Model Uncertainty; Bayesian Model Averaging

What if, when the data arrive, I see that my **initial (prior) parametric** choice for $p(y|\theta)$ was **wrong**?

Having seen the data, I'd now like to **change** $p(y|\theta)$, but **Lindley (1985)** reminds us of **Cromwell's Rule** — if $P(B) = 0$ then $P(B|A) = 0$ for all A, i.e., **anything that has prior probability 0 must have posterior probability 0, no matter how the data come out** — and this appears to say that I **can't change** my initial $p(y|\theta)$ after **looking** at the data.

People have come to refer to $p(y|\theta)$ as the **model**, and to this difficulty as the problem of **model uncertainty**; what now?

I see only three **potential solutions**:

- Start with a **richer set** of **parametric possibilities** $\mathcal{M} = (M_1, M_2, \dots) = (p_1(y|\theta), p_2(y|\theta), \dots)$ in which θ has the same meaning in each model; then Bayes's Theorem becomes

$$p(\theta|y) = \sum_i p(\theta|M_i, y)\, p(M_i|y). \qquad (14)$$

This **(Leamer (1978), Draper (1995))** is **Bayesian model averaging**:

3CV

$$p(\theta|y) = \sum_i p(\theta|M_i, y)\, p(M_i|y)$$

the posterior distribution for θ given the data is a **weighted average** of the **conditional posterior distributions** $p(\theta|M_i, y)$, weighted by the **posterior plausibilities** $p(M_i|y)$ of the models; if $\mathcal{M}$ is **rich enough** I can avoid putting **prior probability 0** on the **actual data behavior**.

- **Looking at the data** to specify the model is a form of **cheating**, but it's okay to cheat **if you pay the right price for doing so**: with **cross-validation methods** such as **3CV (Draper and Krnjajić, 2009)** you **partition** the data into **3 subsets**, use **2** of them for **iterative modeling**, and use the **third subset** to reliably estimate the **predictive accuracy** of the model you arrived at iteratively.

Pretending you know the **actual data-generating mechanism** when you don't (**cheating by looking at the data**) is a form of **understatement of uncertainty**; 3CV solves this by **appropriately widening your uncertainty bands to pay for having cheated**.

Exchangeability; Bayesian Nonparametric Methods

- The other way out is to **not put prior probability 0 on anything**: for example, if I'm about to observe n **continuous** data values $y = (y_1, \dots, y_n)$ on $\mathbb{R}$ (e.g., the lengths of hospital stay of the next n patients with a heart attack diagnosis at the Dominican Hospital in Santa Cruz CA) and I know nothing ahead of time that would **distinguish** one of these patients from another, de Finetti (1930) noticed that my **predictive distribution** $p(y_1, \dots, y_n)$ should be **invariant** to permutation of the labels on the patients; he called such distributions **exchangeable** and proved a remarkable theorem —

Continuous y_i on $\mathbb{R}$ exchangeable → the only **logically consistent** model for the data is

$$\left\{ \begin{array}{ccc} F & \sim & p(F) \\ (y_i|F) & \stackrel{\text{IID}}{\sim} & F \end{array} \right\} \tag{15}$$

where F is the **empirical cumulative distribution function** (CDF) of $(y_1, y_2, \dots)$, $p(F)$ is a probability distribution on the set $\mathcal{F}$ of all CDFs, and IID stands for **independent, identically distributed sampling**.

Placing **probability distributions on functions** involves **Bayesian nonparametric methods** (Ferguson (1973); e.g., Krnajić, Kottas and Draper (2008)): **Pólya trees** and **Dirichlet process priors**.

Laplace; Difficult Integrals Emerge

In theory, at least, the **posterior distribution** $p(\theta|y)$ completely solves the **inference problem** about the unknown θ, 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 θ was a **vector** of real numbers of length k; in this setting no changes are needed to the **notation** —

$$p(\theta|y) = c\, p(\theta)\, l(\theta|y) \tag{16}$$

— 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:

$$\text{with } \theta = (\theta_1, \ldots, \theta_k), \quad c = \left(\int \cdots \int p(\theta)\, l(\theta|y)\, d\theta_1 \cdots d\theta_k \right)^{-1} . \tag{17}$$

The Bayesian Computation Problem

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 $k \geq 10$; moreover, even if you can compute c, for $k > 2$ it's hard to **visualize** a k-dimensional posterior distribution $p(\theta|y)$, so you'll want to look at the k **marginal distributions**

$$p(\theta_j|y) = \int \cdots \int p(\theta|y)\, d\theta_1 \cdots d\theta_{j-1} d\theta_{j+1} \cdots d\theta_k, \tag{18}$$

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**.

The Frequentist Story; Subjectivity and Objectivity

- 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:

The Role of Assumptions and Judgments

(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.

Frequentist Modeling

(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

$$Y_i \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \quad i = 1, \ldots, n \tag{19}$$

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 θ 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 $P(Y_1 = y_1), \ldots, P(Y_n = y_n)$:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n) &= P(Y_1 = y_1) \cdots P(Y_n = y_n) \\ &= \prod_{i=1}^{n} P(Y_i = y_i) \, . \end{aligned} \tag{20}$$

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

The Likelihood Function, Again

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \theta^{y_i} \, (1 - \theta)^{1 - y_i} . \tag{21}$$

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 θ — 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(θ) 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 (21) as a function of θ for fixed y — Fisher called it the **likelihood function** for θ in the Bernoulli(θ) model:

$$\begin{aligned} l(\theta | y) \quad &= \quad l(\theta | y_1, \ldots, y_n) = \prod_{i=1}^{n} \theta^{y_i} \, (1 - \theta)^{1 - y_i} \qquad (22) \\ &= \quad P(Y_1 = y_1, \ldots, Y_n = y_n) \; \text{ but } \textbf{interpreted} \\ & \qquad \textbf{as a function of } \theta \textbf{ for fixed } y. \end{aligned}$$

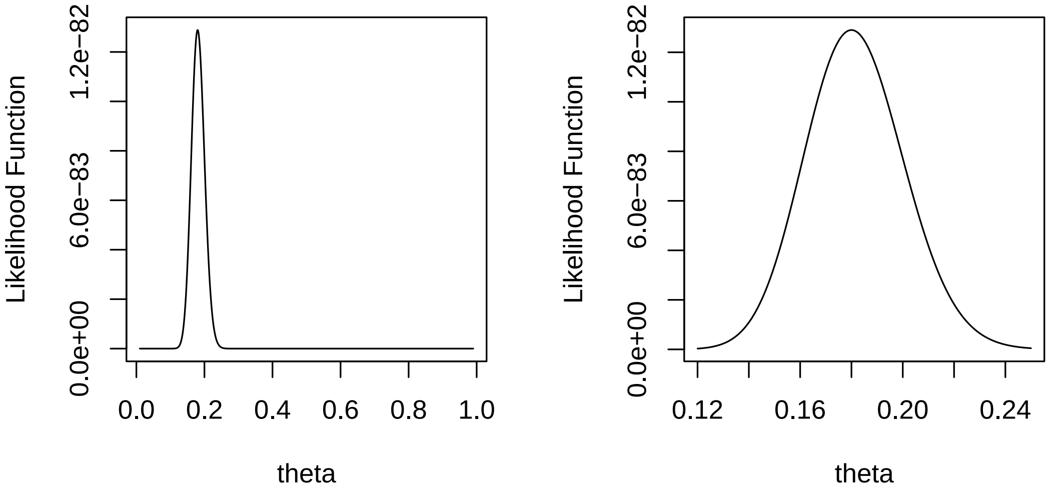
Gaussian Likelihood With Large n

Fisher tried to create a theory of **inference** about θ **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(θ) likelihood function can be **simplified** as follows:

$$l(\theta|y) = \theta^s (1-\theta)^{n-s}, \tag{23}$$

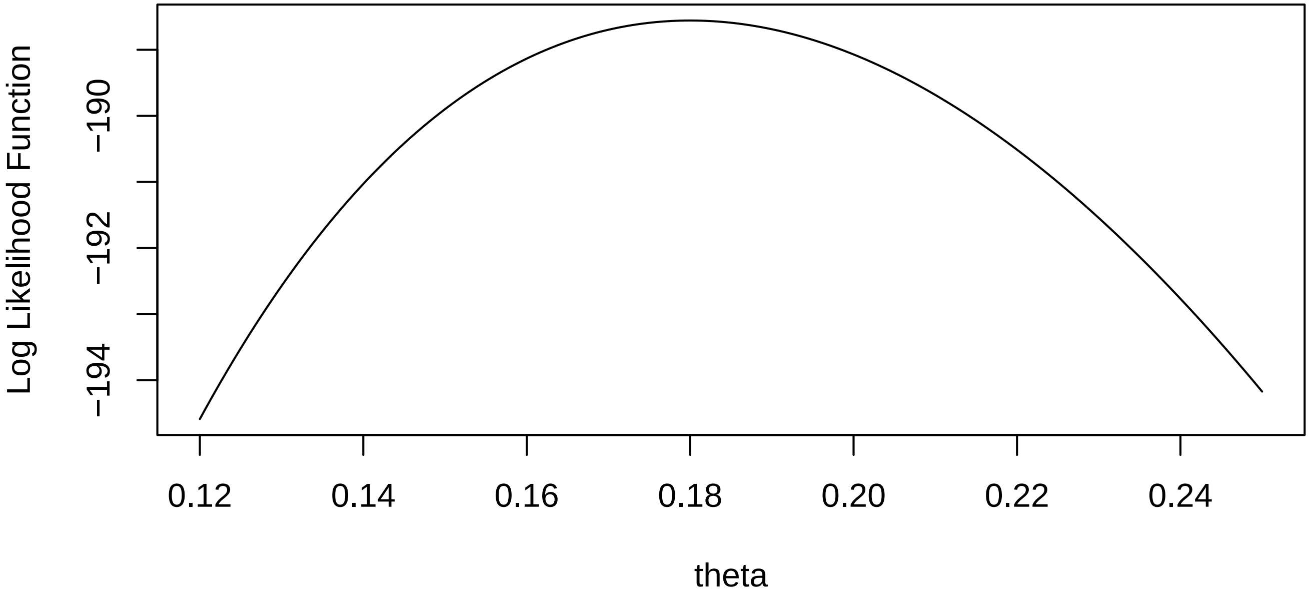
where $s = \sum_{i=1}^n y_i$ is the **number of 1s** in the sample and $(n-s)$ is the **number of 0s**; what does this function **look like**, e.g., with $n = 400$ and $s = 72$ (this is similar to data you would get from the DH: a **30-day mortality rate** from AMI of **18%**)?



Quadratic Log Likelihood With Large 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 θ (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 θ to 0 and solving; here I get the familiar result

$$\hat{\theta}_{\text{MLE}} = \frac{s}{n} = \bar{y}. \tag{24}$$

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

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

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

Frequentist Inference

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** θ of the population of interest (e.g., the **mean**), and I find a reasonable **estimate** $\hat{\theta}$ of θ 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!**) θ (e.g., **confidence intervals** for θ [e.g., "I'm 95% **confident** that θ is between 0.14 and 0.22"] or **hypothesis tests** about θ [e.g., the P **value** for testing H_0: $\theta < 0.1$ against H_A: $\theta \geq 0.1$ is near 0, so I **reject** H_0).

I'm not allowed to make **probability statements** about θ in the **frequentist** paradigm, because θ 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**.

Calibrating the MLE

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 θ 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 $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right)$.

Fisher (1922) also proposed an **approximation** to $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right)$ 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

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

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 θ — in his honor it's now called the (observed) **Fisher information**:

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \left[-\frac{\partial^2}{\partial\theta^2}\log l(\theta|y)\right]_{\theta=\hat{\theta}_{\text{MLE}}} . \tag{26}$$

This quantity **increases** as n goes up, whereas my uncertainty about θ based on the sample, as measured by $\hat{V}\left(\hat{\theta}_{\text{MLE}}\right)$, 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:

$$\hat{V}\left(\hat{\theta}_{\text{MLE}}\right) \doteq \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right) . \tag{27}$$

Likelihood-Based Large-Sample Confidence Intervals

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

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{n}{\hat{\theta}(1-\hat{\theta})} \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{28}$$

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

$$E\left(\hat{\theta}_{\text{MLE}}\right) \doteq \theta, \tag{29}$$

and (b) the sampling distribution of the MLE is approximately **Gaussian** with mean θ and estimated variance given by (27):

$$\hat{\theta}_{\text{MLE}} \overset{\cdot}{\sim} \text{Gaussian}\left[\theta, \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)\right]. \tag{30}$$

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

Repeated-Sampling Asymptotic Optimality of MLE

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

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{n}{\hat{\theta}(1-\hat{\theta})},$$

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 θ 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):

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = O(n) \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = O\left(n^{-1}\right), \tag{31}$$

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 θ 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_{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: $\boxed{P_{B:\text{me}}(Y_1 = y_1, \ldots, Y_n = y_n)}$.

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**:

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

Exchangeability (continued)

$\boxed{\text{NB}}$ **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**.

de Finetti's Theorem For Binary Outcomes

After defining the concept of exchangeability, however, de Finetti went on to prove a **remarkable result**: if I'm willing to regard the $\{Y_i, i = 1, \ldots, n\}$ 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 $\{Y_i, i = 1, \ldots, n\}$, 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 θ) for each of the Y_i must be $P(Y_i = y_i | \theta) = \text{Bernoulli}(y_i | \theta) = \theta^{y_i} (1 - \theta)^{1 - y_i}$,

de Finetti's Theorem (continued)

where P is my **joint probability distribution** on $(Y_1, Y_2, \dots)$;

- $H(t) = \lim_{n \to \infty} P(\bar{Y}_n \le t)$, 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
- $P(Y_1, \dots, Y_n)$ can be expressed as

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \int_0^1 \prod_{i=1}^n \theta^{y_i} (1 - \theta)^{1 - y_i} \, dH(\theta). \tag{32}$$

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), $dH(\theta) = p(\theta) \, d\theta$, (32) can be written in a **more accessible** way as

$$P(Y_1 = y_1, \dots, Y_n = y_n) = \int_0^1 \theta^s (1 - \theta)^{n - s} \, p(\theta) \, d\theta, \tag{33}$$

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 $p(\theta)$,

$$p(y_1, \dots, y_n) = \int_0^1 p(y|\theta) \, p(\theta) \, d\theta = \int_0^1 \theta^s (1-\theta)^{n-s} \, p(\theta) \, d\theta, \qquad (34)$$

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

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

But (a) the left side of (35), interpreted as a function of θ for fixed $y = (y_1, \dots, y_n)$, is recognizable as the **likelihood function** for θ given y, (b) the right side of (35) is recognizable as the likelihood function for θ in **IID Bernoulli sampling**, and (c) (35) 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 $Y_1, \dots, Y_n$ amounts to behaving **as if**

Mixture (Hierarchical) Modeling

- there is a quantity called θ, 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 θ as a **random** quantity with density $p(\theta)$, and
- **conditional** on this θ the Y_i are IID Bernoulli(θ).

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

$$\left\{ \begin{array}{ccc} \theta & \sim & p(\theta) \\ (Y_i|\theta) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta), \ i = 1, \ldots, n \end{array} \right\} . \qquad (36)$$

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(θ)), but a frequentist who treats θ 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 θ 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 θ **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).

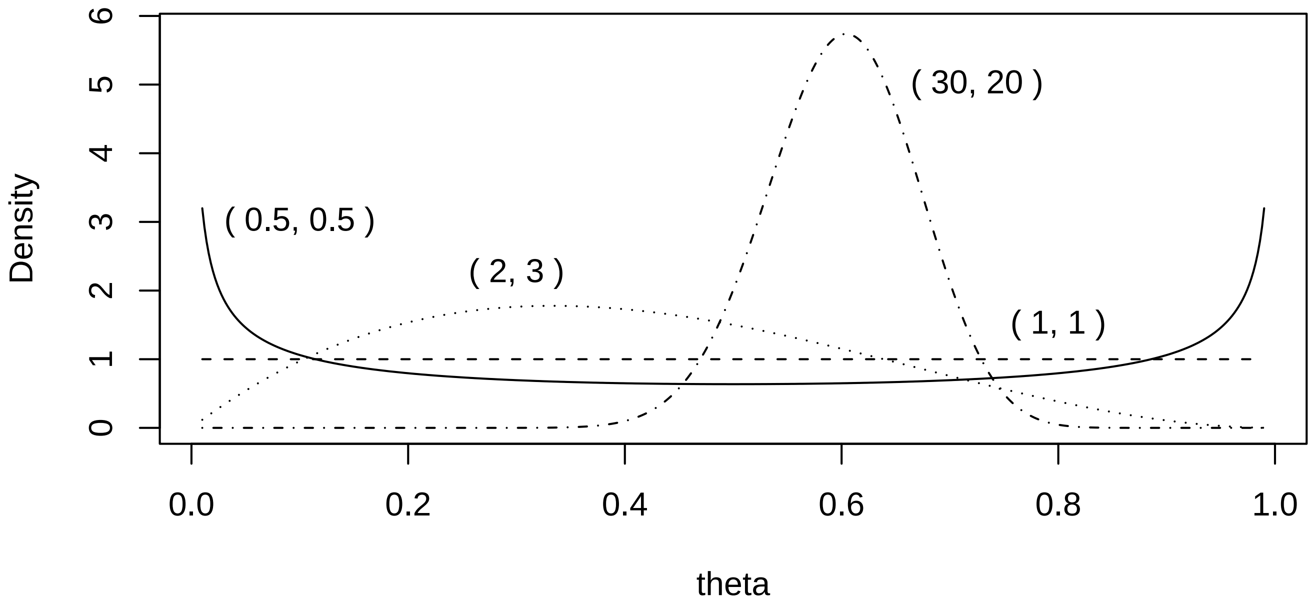
The Beta Family of Densities on $(0, 1)$

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,

$$\text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad (37)$$

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,

$$p(y_1, \ldots, y_n|\theta) = l(\theta|y) = \theta^s (1 - \theta)^{n-s}, \tag{38}$$

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)

$\boxed{\text{Conjugate analysis}}$ — 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 θ 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

$$E(\theta) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad V(\theta) = \frac{\alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)}. \tag{39}$$

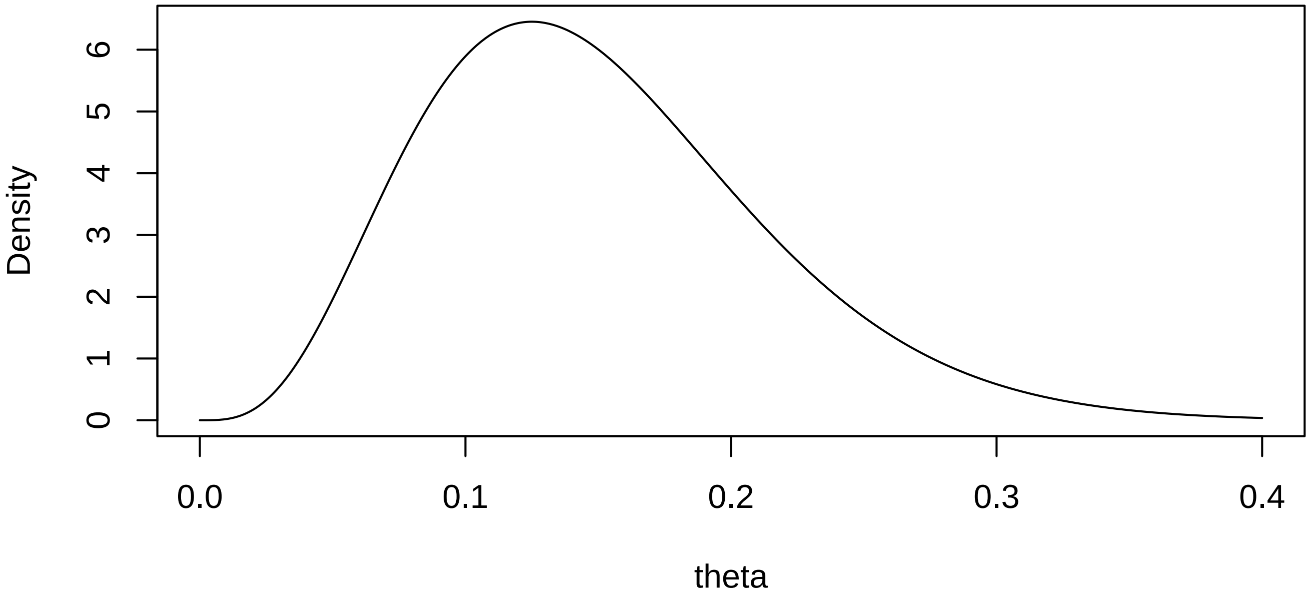
Setting $\frac{\alpha}{\alpha+\beta} = 0.15$ and **solving** for β yields $\beta = \frac{17}{3}\alpha$; then the equation

$$0.95 = \int_{0.05}^{0.30} \text{Beta}\left(\theta \,\middle|\, \alpha, \frac{17}{3}\alpha \right) d\theta \tag{40}$$

can readily be **solved numerically** for α (e.g., in a **symbolic computing**

The Beta Family (continued)

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 (α, β) 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 α and β are called $\boxed{\textbf{hyperparameters}}$ since they're parameters of the prior distribution.

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

$$\begin{array}{rcll} (\alpha, \beta) & = & (4.5, 25.5) & \text{(hyperparameters)} \\ (\theta | \alpha, \beta) & \sim & \text{Beta}(\alpha, \beta) & \text{(prior)} \\ (Y_1, \dots, Y_n | \theta) & \stackrel{\text{IID}}{\sim} & \text{Bernoulli}(\theta) & \text{(likelihood)} \end{array} \qquad (41)$$

(41) 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 (α, β) 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 $Y_i, i = 1, \ldots, n$ and s as the sum of the y_i (a **sufficient statistic** for θ, as noted above, with the Bernoulli likelihood),

$$\begin{aligned} p(\theta|y, \alpha, \beta) &= c\, l(\theta|y)\, p(\theta|\alpha, \beta) \\ &= c\, \theta^s\, (1-\theta)^{n-s}\, \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= c\, \theta^{(s+\alpha)-1} (1-\theta)^{(n-s+\beta)-1}, \end{aligned} \tag{42}$$

i.e., the **posterior** for θ 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 (Beta(α, β)) 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**.

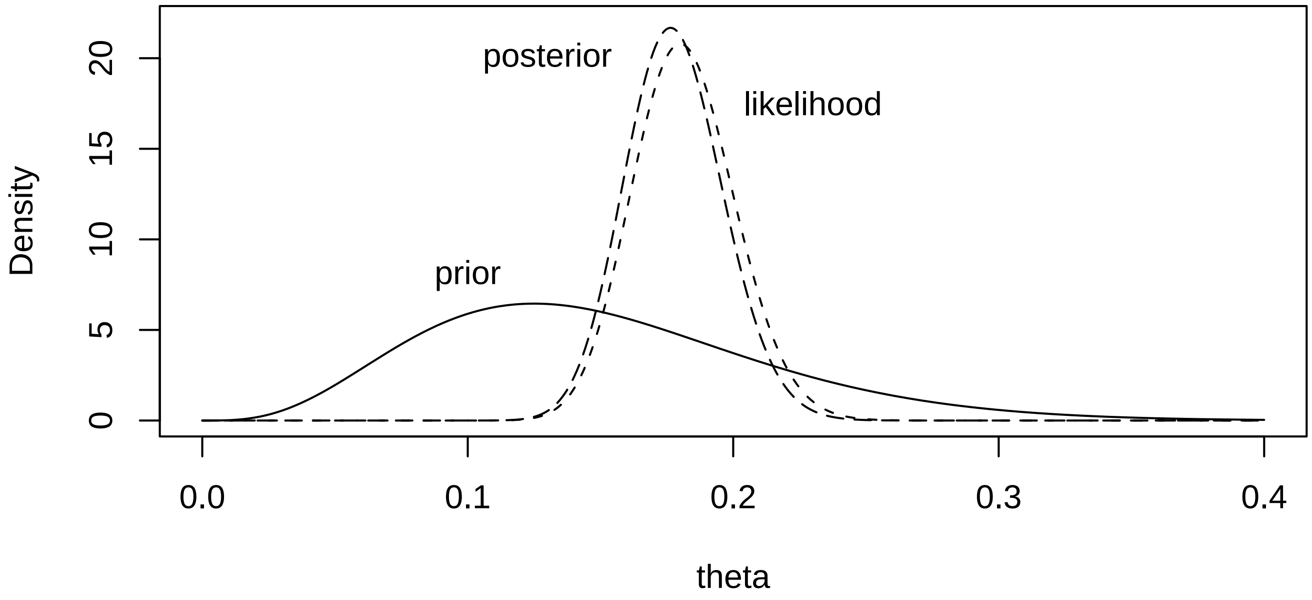
Prior-To-Posterior Updating

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

$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (Y_i | \theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\theta | y) \sim \text{Beta}(\alpha + s, \beta + n - s), \qquad (43)$$

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

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



Prior-To-Posterior Updating (continued)

Then the **prior** is Beta(4.5, 25.5), the **likelihood** is Beta(73, 329), the **posterior** for θ 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(α, β) 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:

$$
\begin{array}{ccccccccc}
\dfrac{\alpha+s}{\alpha+\beta+n} & = & \left(\dfrac{\alpha+\beta}{\alpha+\beta+n}\right) & \left(\dfrac{\alpha}{\alpha+\beta}\right) & + & \left(\dfrac{n}{\alpha+\beta+n}\right) & \left(\dfrac{s}{n}\right) \\
\begin{array}{c}\text{posterior}\\ \text{mean}\end{array} & = & \left(\begin{array}{c}\text{prior}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{prior}\\ \text{mean}\end{array}\right) & + & \left(\begin{array}{c}\text{data}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{data}\\ \text{mean}\end{array}\right) \\
.178 & = & (.070) & (.15) & + & (.93) & (.18)
\end{array}
$$

Comparison With Frequentist Modeling

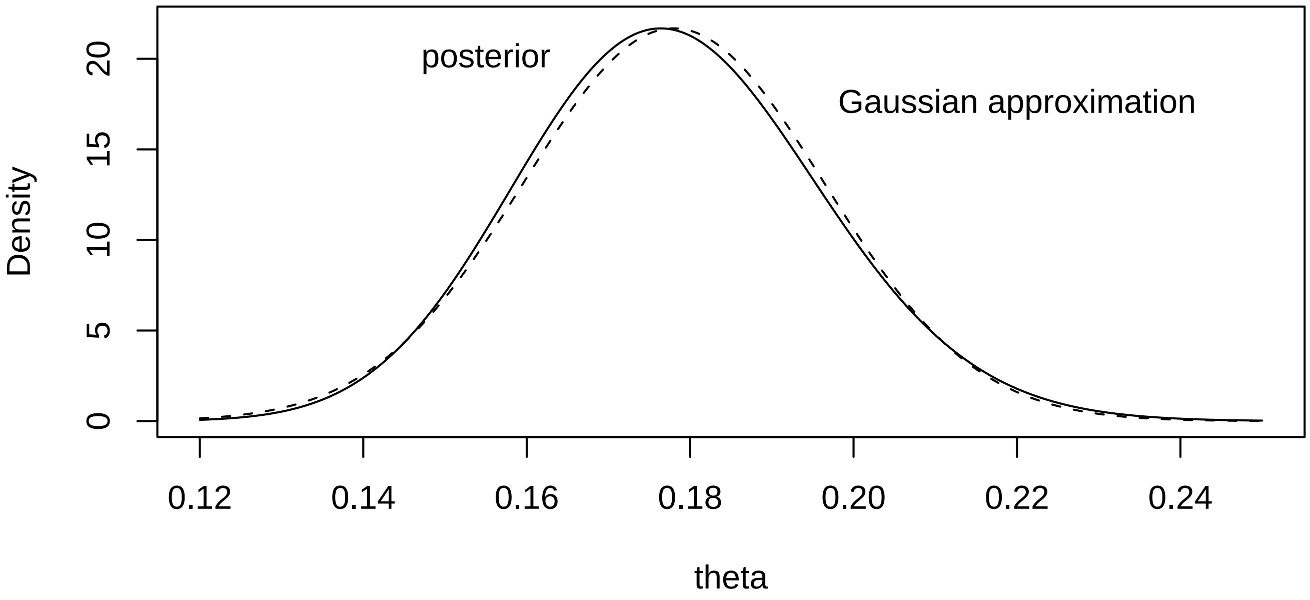
Another way to put this is that the data mean, $\bar{y} = \frac{s}{n} = \frac{72}{400} = .18$, has been **shrunk** toward the prior mean .15 by (in this case) a modest amount: the posterior mean is about .178, and the **shrinkage factor** is $\frac{30}{30+400} =$ about .07.

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 $\text{N}\left[\theta, \frac{\theta(1-\theta)}{n}\right]$, so an approximate **95% confidence interval** for θ 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 θ 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 α^* and β^* are the parameters of the beta posterior distribution; a **95% central posterior interval** for θ 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 θ **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 $(\bar{Y} - \theta) \mathrel{\dot{\sim}} \text{Gaussian}(0, .019^2)$; whereas

Comparison With Frequentist Modeling (continued)

- In the Bayesian approach $\bar{y}$ **is fixed at its observed value and** θ **is treated as random**, as a means of quantifying my posterior uncertainty about it: $(\theta - \bar{y}|\bar{y}) \mathrel{\dot{\sim}} \text{Gaussian}(0, .018^2)$.

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

$\boxed{\textbf{Q:}}$ When does **maximum likelihood** work well from a **Bayesian** viewpoint?

$\boxed{\textbf{A:}}$ (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**;

Testing; Bayesian Decision Theory

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: $\theta \geq 0.1$) 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

Gibbs Sampling; Bayesian Applied Statistics

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**:

- θ, something of **interest** which is **unknown** (or only partially known) to me (e.g., θ_{RR}, the relative risk of getting a disease under one treatment condition versus another).

Often θ 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 θ.

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 θ 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 θ, **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:

$$\begin{aligned} p(\theta|y, \mathcal{B}) &= c\, p(\theta|\mathcal{B})\, l(\theta|y, \mathcal{B}) \\ p(y^*|y, \mathcal{B}) &= \int p(y^*|\theta, \mathcal{B})\, p(\theta|y, \mathcal{B})\, d\theta \\ a^* &= \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y, \mathcal{B})} \left[U(a, \theta) \right] \end{aligned} \tag{44}$$

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

$\mathcal{B}$ is often **omitted** from the basic equations (sometimes with **unfortunate consequences**), yielding the **simpler-looking** forms

$$\begin{gathered} p(\theta|y) = c\, p(\theta)\, l(\theta|y) \qquad p(y^*|y) = \int p(y^*|\theta)\, p(\theta|y)\, d\theta \\ a^* = \operatorname*{argmax}_{a \in \mathcal{A}} E_{(\theta|y)} \left[U(a, \theta) \right] \end{gathered} \tag{45}$$

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 θ given $\mathcal{B}$, in the form of a **probability density function** (PDF) or **probability mass function** (PMF) if θ 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 θ given y and $\mathcal{B}$, which summarizes my **current total information** about θ and **solves** the **inference problem**.

These are actually **not very good names** for $p(\theta|\mathcal{B})$ and $p(\theta|y,\mathcal{B})$, because (e.g.) $p(\theta|\mathcal{B})$ really stands for **all (relevant) information** about θ (given $\mathcal{B}$) **external to** y, whether that information was obtained **before** (or **after**) y arrives, 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

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

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

- $p(y^*|\theta,\mathcal{B})$ is my sampling distribution for **future data values** y^* given θ 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).

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

$$l(\theta|y,\mathcal{B}) = c\,p(y|\theta,\mathcal{B}). \tag{46}$$

- $\mathcal{A}$ 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 θ, and the third equation says I should **find** the action a^* that **maximizes expected utility** (MEU).

Predictive Distributions and MCMC

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

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

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

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).

An Example of Poorly-Calibrated Frequentist Inference

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:

$$\begin{array}{c} y_{ij} = \beta_0 + u_j + e_{ij}, \quad i = 1, \ldots, n_j, \quad j = 1, \ldots, J; \\ \sum_{j=1}^{J} n_j = N, \quad (u_j | \sigma_u^2) \stackrel{\text{IID}}{\sim} N(0, \sigma_u^2), \quad (e_{ij} | \sigma_e^2) \stackrel{\text{IID}}{\sim} N(0, \sigma_e^2). \end{array} \tag{47}$$

Browne and Draper (2006) used a **simulation study** to show that, with a variety of **maximum-likelihood-based methods** for creating **confidence intervals** for σ_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

$$\begin{array}{c} (y_{ijk} \,|\, p_{ijk}) \stackrel{\text{indep}}{\sim} \text{Bernoulli}(p_{ijk}) \qquad \text{with} \\ \text{logit}(p_{ijk}) = \beta_0 + \beta_1 x_{1ijk} + \beta_2 x_{2jk} + \beta_3 x_{3k} + u_{jk} + v_k , \end{array} \qquad (48)$$

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 σ_u^2 and σ_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.

Frequentist Hypothesis Tests

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

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

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

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

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

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

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

(subject to the **constraint** P_F(type I error) $\leq \alpha$), in which β is typically some **conventional value** like **0.2** ($1 - \beta =$ **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 Volunteers	241/3598 (6.7)	127/1805 (7.0)	6 (–17 to 24)	.59	> .5	
Black (Non-Hisp)	6/233 (2.6)	9/116 (7.8)	67 (6 to 88)	.028	.24	
Black Women	1/112 (0.9)	4/57 (7.0)	87 (19 to 98)	.033		
Nonwhite	30/604 (5.0)	29/310 (9.4)	47 (12 to 68)	.012	.13	
Nonwhite Men	27/461 (6.1)	25/236 (10.6)	43 (3 to 67)	.036		

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 (%) of HIV–1 Infection		VE	*P* Value		
Group	Vaccine	Placebo	(95% CI)	Unadj	Adj	D-M
Nonwhite	30/604 (5.0)	29/310 (9.4)	47 (12 to 68)	.012	.13	

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

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 Δ 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 $a_1 = \{$**implement** $T\}$ and $a_2 = \{$**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 Δ based on **clinical trial data** D with **sample size** n and Δ^* is chosen so that the corresponding P value is no greater than α, the **type I error probability** (the chance of **rejecting** H_0 when H_0 is **true**).

Details (continued)

As noted above, α 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** β 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 α and β, 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 Δ, but I can look at an **oversimplified utility structure** that permits **comparison with hypothesis-testing**: for $u_{ij} \geq 0$,

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

Details (continued)

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

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 (α) corresponding to **one** of these **four ingredients** ($-u_{12}$), and even then α 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

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

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 P **values** that are **much larger** than the usual threshold (e.g., **0.3** instead of **0.05**).