# Bayesian Modeling, Inference, Prediction and Decision-Making

## 2: Exchangeability and Conjugate Modeling

David Draper

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

Short Course (Days 1 and 2)
University of Reading (UK)

## 2.1 Probability as quantification of uncertainty about observables; binary outcomes

**Case Study:** *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.

# The Meaning of Probability

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?

**The meaning of probability.** I'm definitely **uncertain** about the 0–1 death outcomes $Y_1, \ldots, Y_n$ before I observe any of them.

**Probability** is supposed to be the part of mathematics concerned with quantifying uncertainty; can probability be used here?

In part 1 I argued that the answer was **yes**, and that three types of probability — **classical**, **frequentist**, and **Bayesian** — are available (in principle) to quantify uncertainty like that encountered here.

The **classical** approach turns out to be **impractical** to implement in all but

the simplest problems; I'll focus here on the **frequentist** and **Bayesian** stories.

**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.

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

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

# What is the Population?

for some $0 < \theta < 1$, but what's the **population** to which it's appropriate to **generalize outward** from the 400 1s and 0s that will be observed?

Here are some **possibilities**:

- (Fisher) All AMI patients who **might have** come to the DH in 2006–09 if the world had turned out differently; or
- Assuming sufficient **time-homogeneity** in all relevant factors, I could try to argue that the collection of all 400 AMI patients at the DH from 2006–09 is **like** a random sample of size 400 from the population of all AMI patients at the DH from (say) 2000–2015; or
- **Cluster sampling** is a way to choose, e.g., patients by taking a random sample of hospitals and then a random sample of patients **nested** within those hospitals; what we actually have here is a kind of cluster sample of **all** 400 AMI patients from the DH in 2006–2009.

  Cluster samples tend to be less informative than simple random samples (SRSs) of the same size because of (positive) **intracluster correlation** (patients in a given hospital tend to be more similar in their outcomes than

# Frequentist Modeling (continued)

would an SRS of the same size from the population of all the patients in all the hospitals).

Assuming the DH to be representative of some broader collection of hospitals in California and (unwisely) ignoring intracluster correlation, I could try to argue that these 400 1s and 0s were **like** a simple random sample of 400 AMI patients from this larger collection of hospitals.

None of these options is entirely **compelling**.

If I'm willing to pretend the data are like a sample from some population, interest would then focus on inference about the **parameter** $\theta$, the "underlying death rate" in this larger collection of patients to which I feel comfortable generalizing the 400 1s and 0s: if $\theta$ were unusually high, that would be **prima facie** evidence of a possible quality of care problem.

Suppose (**as above**) that the frequentist model is

$$Y_i \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \quad i = 1, \ldots, n \quad \text{for some } 0 < \theta < 1. \tag{2}$$

Since the $Y_i$ are **independent**, the **joint** sampling distribution of all of them,

# Frequentist Modeling (continued)

$P(Y_1 = y_1, \ldots, Y_n = y_n)$, is the **product** of the separate, or **marginal**, sampling distributions $P(Y_1 = y_1), \ldots, P(Y_n = y_n)$:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n) &= P(Y_1 = y_1) \cdots P(Y_n = y_n) \\ &= \prod_{i=1}^{n} P(Y_i = y_i)\,. \end{aligned} \tag{3}$$

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

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \prod_{i=1}^{n} \theta^{y_i}\,(1-\theta)^{1-y_i}\,. \tag{4}$$

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.

# The Likelihood Function

In 1921 (as you know) Fisher had the following idea (Laplace (1774) had it first): **after** the data have arrived it makes more sense to interpret (4) as a function of $\theta$ for fixed $y$ — this is the **likelihood function** for $\theta$ in the Bernoulli($\theta$) model:

$$\begin{aligned} l(\theta|y) &= l(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \\ &= P(Y_1 = y_1, \ldots, Y_n = y_n) \text{ but } \mathbf{interpreted} \\ &\quad \mathbf{as\ a\ function\ of\ } \theta \mathbf{\ for\ fixed\ } y. \end{aligned} \tag{5}$$

Fisher tried to create a theory of **inference** about $\theta$ **based only on this function** — this turns out to be an important ingredient, **but not the only important ingredient**, in inference from the Bayesian viewpoint.

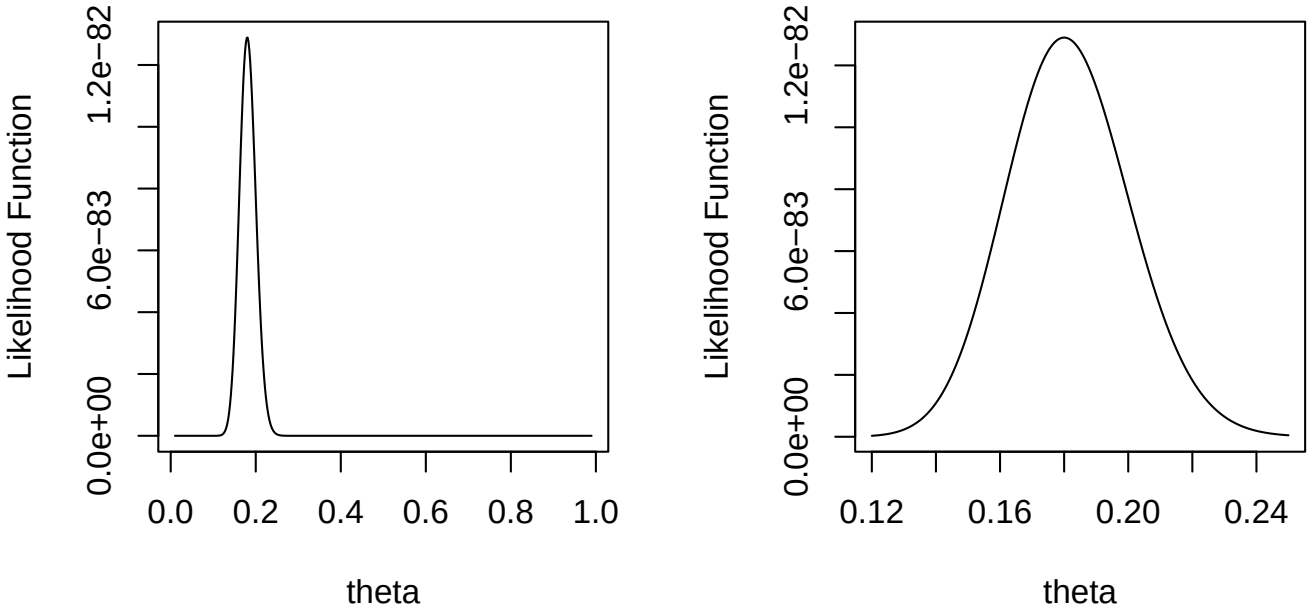
The Bernoulli($\theta$) likelihood function can be **simplified** as follows:

$$l(\theta|y) = \theta^s (1-\theta)^{n-s}, \tag{6}$$

where $s = \sum_{i=1}^{n} y_i$ is the **number of 1s** in the sample and $(n-s)$ is the **number of 0s**.

# The Likelihood Function (continued)

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%**)?

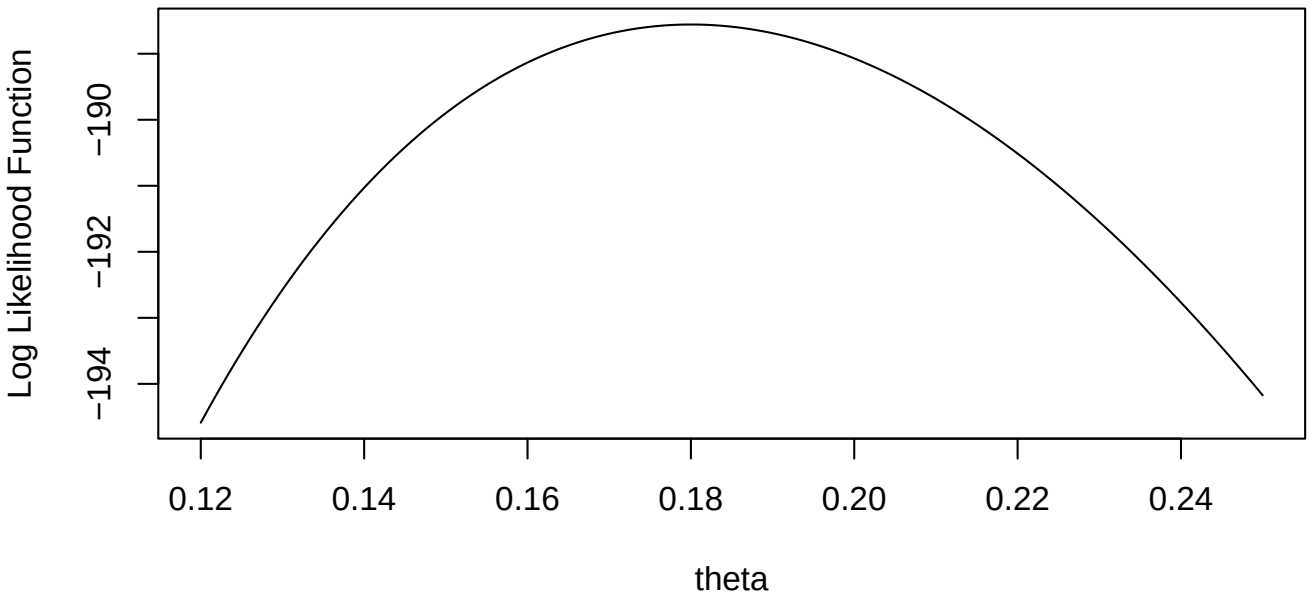



This looks a lot like a **Gaussian distribution** (not yet density-normalized) for $\theta$, which is the **Bayesian** way to **interpret** the likelihood function (see below).

# Likelihood and Log Likelihood

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 $s = \sum_{i=1}^{n} y_i$** — (as you know) 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**.

# Maximizing the Likelihood Function

Fisher had (as you know) the further 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

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

The function of the data that maximizes the likelihood (or log likelihood) function is (as you know) 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**) $c > 0$.

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

Fisher (1922) also proposed (as you know) 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

$$V\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\theta(1-\theta)}{n} \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{7}$$

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

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \left[-\frac{\partial^2}{\partial\theta^2}\log l(\theta|y)\right]_{\theta=\hat{\theta}_{\text{MLE}}}. \tag{8}$$

This quantity **increases** as $n$ goes up, whereas my uncertainty about $\theta$ 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:

$$\hat{V}\left(\hat{\theta}_{\text{MLE}}\right) \doteq \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right). \tag{9}$$

# Likelihood-Based Large-Sample Confidence Intervals

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

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{n}{\hat{\theta}(1-\hat{\theta})} \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = \frac{\hat{\theta}(1-\hat{\theta})}{n}, \tag{10}$$

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

$$E\left(\hat{\theta}_{\text{MLE}}\right) \doteq \theta, \tag{11}$$

and (b) the sampling distribution of the MLE is approximately **Gaussian** with mean $\theta$ and estimated variance given by (9):

$$\hat{\theta}_{\text{MLE}} \dot{\sim} \text{Gaussian}\left[\theta, \hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)\right]. \tag{12}$$

Thus for large $n$ an **approximate 95% confidence interval** for $\theta$ is given by

$$\hat{\theta}_{\text{MLE}} \pm 1.96\sqrt{\hat{I}^{-1}\left(\hat{\theta}_{\text{MLE}}\right)}.$$

# 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 $\theta$ 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):

$$\hat{I}\left(\hat{\theta}_{\text{MLE}}\right) = O(n) \quad \text{and} \quad \hat{V}\left(\hat{\theta}_{\text{MLE}}\right) = O\left(n^{-1}\right), \tag{13}$$

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

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**.

I'll argue later 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.

## 2.3 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)

**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 $\theta$) for each of the $Y_i$ must be $P(Y_i = y_i|\theta) = \text{Bernoulli}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$,

# de Finetti's Theorem (continued)

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

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

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \int_0^1 \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, dH(\theta). \tag{14}$$

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), $dH(\theta) = p(\theta)\, d\theta$, (14) can be written in a more accessible way as

$$P(Y_1 = y_1, \ldots, Y_n = y_n) = \int_0^1 \theta^s (1-\theta)^{n-s} \, p(\theta) \, d\theta, \tag{15}$$

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

# Generalizing Outward from the Observables

**Important digression 1.** Some **awkwardness** arose above in the **frequentist** approach to modeling the AMI mortality data, because it was not clear what population $\mathcal{P}$ the data could be regarded as **like a random sample from**.

This awkwardness also arises in **Bayesian** modeling: even though in practice I'm only going to observe $(y_1, \ldots, y_n)$, de Finetti's representation theorem requires me to **extend my judgment of finite exchangeability** to the **countably-infinite collective** $(y_1, y_2, \ldots)$,

$\rightarrow$ **and this is precisely like viewing** $(y_1, \ldots, y_n)$ **as a random sample from** $\mathcal{P} = (y_1, y_2, \ldots)$.

The key point is that the **difficulty** arising from **lack of clarity** about the **scope of valid generalizability** from a given set of observational data is a **fundamental scientific problem** that emerges whenever purely **observational** data are viewed through an **inferential** or **predictive** lens, whether the statistical methods I use are **frequentist** or **Bayesian**.

# The Law of Total Probability

**Important digression 2.** It's a general fact (as you know) about **true-false propositions** $D$ and $A$ that

$$\begin{aligned} P(D) &= P(D \text{ and } A) + P[D \text{ and (not } A)] \\ &= P(A)\,P(D|A) + P(\text{not } A)\,P(D|\text{not } A). \end{aligned} \tag{16}$$

This is a special case of the **Law of Total Probability** (LTP).

$A$ and (not $A$) divide, or **partition**, the collection of all possible outcomes into two non-overlapping (**mutually exclusive**) and **exhaustive** possibilities.

Let $A_1, \ldots, A_k$ be any **finite partition**, i.e., $P(A_i \text{ and } A_j) = 0$ (mutually exclusive) and $\sum_{i=1}^{k} P(A_i) = 1$ (exhaustive); then a **more general** version of the LTP gives that

$$\begin{aligned} P(D) &= P(D \text{ and } A_1) + \ldots + P(D \text{ and } A_k) \\ &= P(A_1)\,P(D|A_1) + \ldots + P(A_k)\,P(D|A_k) \\ &= \sum_{i=1}^{k} P(A_i)\,P(D|A_i). \end{aligned} \tag{17}$$

# Hierarchical (Mixture) Modeling

There is (as you know) a **continuous** version of the LTP: by analogy with (17), if $X$ and $Y$ are real-valued random variables

$$p(y) = \int_{-\infty}^{\infty} p(x)\, p(y|x)\, dx. \tag{18}$$

$p(x)$ in this expression can be thought of as a **mixing distribution**.

Intuitively (18) says that the overall probability behavior $p(y)$ of $Y$ is a mixture (**weighted average**) of the conditional behavior $p(y|x)$ of $Y$ given $X$, weighted by the behavior $p(x)$ of $X$.

Another way to put this is to say that I have a choice: I can either model the random behavior of $Y$ **directly**, through $p(y)$, or **hierarchically**, by first modeling the random behavior of $X$, through $p(x)$, and then modeling the conditional behavior of $Y$ given $X$, through $p(y|x)$.

Notice that $X$ and $Y$ are **completely general** in this discussion — in other words, given any quantity $Y$ that I want to model stochastically, I'm free to choose any $X$ (upon which $Y$ depends) and model $Y$ **hierarchically** given $X$ instead, if that's easier.

# Hierarchical (Mixture) Modeling (continued)

**Symbolically**

$$Y \quad \leftrightarrow \quad \left\{ \begin{array}{c} X \\ Y|X \end{array} \right\}. \tag{19}$$

The reason for bringing all of this up now is that (15) can be **interpreted** as follows, with $\theta$ playing the role of $x$:

$$\begin{aligned} p(y_1, \ldots, y_n) &= \int_0^1 p(y_1, \ldots, y_n|\theta)\, p(\theta)\, d\theta \\ &= \int_0^1 \theta^s (1-\theta)^{n-s}\, p(\theta)\, d\theta. \end{aligned} \tag{20}$$

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

$$p(y_1, \ldots, y_n|\theta) = \theta^s (1-\theta)^{n-s}. \tag{21}$$

But (a) the left side of (21), 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 (21) is recognizable as the likelihood function for $\theta$ in **IID Bernoulli sampling**, and (c) (21) 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**

- 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

$$\left\{ \begin{array}{ccc} \theta & \sim & p(\theta) \\ (Y_i|\theta) & \overset{\text{IID}}{\sim} & \text{Bernoulli}(\theta),\ i = 1, \ldots, n \end{array} \right\}. \tag{22}$$

# Exchangeability and Conditional Independence

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** $p(\theta)$.

**NB** This is the first example of a general fact:

$$Y_i \text{ exchangeable} \leftrightarrow \left\{ \begin{array}{c} Y_i \textbf{ conditionally IID} \\ \text{given one or more parameters} \end{array} \right\}. \tag{23}$$

So **exchangeability** is a special kind of **conditional independence**: binary exchangeable $y_i$ are not independent, but they become conditionally independent given $\theta$.

(22) is an example of the simplest kind of **hierarchical model (HM)**: a model at the top level for the underlying death rate $\theta$, and then a model below that for the binary mortality indicators $Y_i$ conditional on $\theta$ (this is a basic instance of (19): it's **not easy** to model the **predictive** distribution for $(Y_1, \ldots, Y_n)$ directly, but it becomes a lot easier when $\theta$ is introduced at the **top level of a 2–level hierarchy**).

# Mixing Distribution = Prior Distribution

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 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.

| Approach | Fixed | Random |
|---|---|---|
| **Frequentist** | $\theta$ | $Y$ |
| **Bayesian** | $y$ | $\theta$ |

## 2.4 Prior, posterior, and predictive distributions

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 $y$ — as noted in Part 1, it has become traditional to call it my **prior distribution** for $\theta$ (I'll address how one might go about **specifying** this distribution below).

# Bayes' Theorem

**Q:** If $p(\theta)$ represents my information external to $\theta$, what represents this information **after** $y$ has been observed?

**A:** It has to be $p(\theta|y)$, the **conditional** distribution for $\theta$ given how $y$ came out.

It's **conventional** (again appealing to terms involving time) to call this the **posterior distribution** for $\theta$ given $y$.

**Q:** How do I get from $p(\theta)$ to $p(\theta|y)$, i.e., how do I **update** my information about $\theta$ in light of the data?

**A:** **Bayes' Theorem** for **continuous** quantities:

$$p(\theta|y) = \frac{p(\theta)\, p(y|\theta)}{p(y)}. \tag{24}$$

This requires some interpreting. As a Bayesian I'm **conditioning on the data**, i.e., I'm thinking of the left-hand side of (24) as a function of $\theta$ for fixed $y$, so that must also be true of the right-hand side. Thus

(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;

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

We've already encountered this idea (page 8): $l(\theta|y) = c\,p(y|\theta)$ is the **likelihood function**.

So **Bayes' Theorem** becomes

$$p(\theta|y) = c \cdot p(\theta) \cdot l(\theta|y) \tag{25}$$

$$\text{posterior} = \left(\begin{array}{c}\text{normalizing}\\ \text{constant}\end{array}\right) \cdot \text{prior} \cdot \text{likelihood}\ .$$

You can also readily construct **predictive distributions** for the $y_i$ before they're observed, or for future $y_i$ once some of them are known.

For example, by the LTP, the **posterior predictive distribution** for $(y_{m+1}, \ldots, y_n)$ given $(y_1, \ldots, y_m)$ is

# Predictive Distributions (continued)

$$p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m) = \tag{26}$$
$$\int_0^1 p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)\, p(\theta | y_1, \ldots, y_m)\, d\theta.$$

Consider $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$: if I **knew** $\theta$, the information $y_1, \ldots, y_m$ about how the first $m$ of the $y_i$ came out would be **irrelevant** (imagine predicting the results of IID coin-tossing: if I somehow **knew** that the coin was perfectly fair, i.e., that $\theta = 0.5$, then getting (say) 6 heads in the first 10 tosses would be useless to me in quantifying the likely behavior of the next (say) 20 tosses — I'd just use the **known true value** of $\theta$).

Thus $p(y_{m+1}, \ldots, y_n | \theta, y_1, \ldots, y_m)$ is just $p(y_{m+1}, \ldots, y_n | \theta)$, which in turn is just the **sampling distribution** under IID B($\theta$) sampling for the binary observables $y_{m+1}, \ldots, y_n$, namely $\prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$.

And finally $p(\theta | y_1, \ldots, y_m)$ is recognizable as just the **posterior distribution** for $\theta$ given the first $m$ of the binary outcomes.

**Putting this all together** gives

# Parameters and Observables

$$p(y_{m+1}, \ldots, y_n | y_1, \ldots, y_m) = \tag{27}$$
$$= \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, p(\theta | y_1, \ldots, y_m) \, d\theta$$

(I can't compute (27) yet because $p(\theta | y_1, \ldots, y_m)$ depends on $p(\theta)$, which I haven't **specified** so far).

This also brings up a key difference between a **parameter** like $\theta$ on the one hand and the $Y_i$, before I've observed any data, on the other: parameters are inherently **unobservable**.

This makes it harder to evaluate the **quality** of my uncertainty assessments about $\theta$ than to do so about the **observable** $y_i$: to see how well I'm doing in predicting observables I can just compare my predictive distributions for them with how they actually turn out, but of course this isn't possible with parameters like $\theta$ **which I'll never actually see**.

The de Finetti approach to modeling emphasizes the **prediction** of observables as a valuable adjunct to **inference** about unobservable parameters, for at least two reasons:

# The Value of Predictive Thinking

- Key scientific questions are often **predictive** in nature: e.g., rather than asking "Is drug A better than B (on average across many patients) for lowering blood pressure?" (inference) the ultimate question is "How much more will drug A lower **this patient's** blood pressure than drug B?" (prediction); and
- Good **diagnostic checking** is predictive: An inference about an unobservable parameter can never be directly verified, but often I can reasonably conclude that inferences about the parameters of a model which produces poor predictions of observables are also **suspect**.

With the predictive approach parameters diminish in importance, especially those that have no physical meaning — such parameters (unlike $\theta$ above) are just **place-holders for a particular kind of uncertainty on my way to making good predictions**.

It's arguable (e.g., Draper, 1995) that the discipline of statistics, and particularly its applications in the social sciences, would be improved by a **greater emphasis on predictive feedback**.

# Where Does the Prior Come From?

This is not to say that parametric thinking should be **abolished**.

As the calculations on the previous pages emphasized, parameters play an important simplifying role in forming modeling judgments: the single strongest simplifier of a joint distribution is **independence** of its components, and whereas, e.g., in the mortality example the $Y_i$ are not themselves independent, they become so conditional on $\theta$.

---

de Finetti's Theorem for 0–1 outcomes says informally that if I'm trying to make **coherent** (internally consistent) probability judgments about a series of 1s and 0s that I judge exchangeable, I may as well behave like a frequentist — IID Bernoulli($\theta$) — with a prior distribution $p(\theta)$; but **where does the prior come from**?

**NB** Coherence doesn't help in answering this question — it turns out that **any** prior $p(\theta)$ could be part of **somebody**'s coherent probability judgments.

Some people regard the need to answer this question in the Bayesian approach as a **drawback**, but it seems to me to be a **positive feature**, as follows.

# Predictive Calibration

From Bayes' Theorem the prior is supposed to be a summary of what I know (and don't know) about $\theta$ external to the data set $(y_1, \ldots, y_n)$: from previous datasets of which I'm aware, from the relevant literature, from expert opinion, ... from all "good" source(s), if any exist.

**Such information is almost always present, and should presumably be used when available; the issue is how to do so "well."**

> The goal is evidently to choose a prior that I'll **retrospectively** be **proud of**, in the sense that my predictive distributions for the observables (a) are well-centered near the actual values and (b) have uncertainty bands that correspond well to the realized discrepancies between actual and predicted values; this is a form of calibration of my probability judgments.

There is **no guaranteed way to do this**, just as there is no guaranteed way to arrive at a "good" frequentist model (see "Where does the likelihood come from?" below).

Some general comments on arriving at a **"good"** prior:

# Choosing a "Good" Prior

- There is a growing literature on methodology for **elicitation** of prior information (e.g., Kadane et al., 1980; Craig et al., 1997; Kadane and Wolfson, 1997; O'Hagan, 1997), which brings together ideas from statistics and perceptual psychology (e.g., people turn out to be better at estimating **percentiles** of a distribution than they are at estimating **standard deviations** (SDs)).
- Bayes' Theorem on the **log scale** says (apart from the normalizing constant)

$$\log(\text{posterior}) = \log(\text{prior}) + \log(\text{likelihood}), \tag{28}$$

  i.e., (posterior information) = (data information) + (prior information). This means that **close attention should be paid to the information content of the prior** by, e.g., density-normalizing the likelihood and plotting it on the same scale as the prior: it's possible for small $n$ for the **prior to swamp the data**, and in general I shouldn't let this happen without a good reason for doing so.

  Comfort can also be taken from the other side of this coin: with large $n$ (in

# Prior Specification (continued)

many situations, at least) the **data will swamp the prior**, and specification errors become less important.

- When I notice I'm quite uncertain about how to specify the prior, I can try **sensitivity** or **pre-posterior analysis**: exploring the mapping from prior to posterior, before the data are gathered, by (a) generating some possible values for the observables, (b) writing down several plausible forms for the prior, and (c) carrying these forward to posterior distributions — if the resulting distributions are similar (i.e., if "all reasonable roads lead to Rome"), I've uncovered a useful form of stability in my results; if not I can try to capture the prior uncertainty **hierarchically**, by, e.g., adding another layer to a model like (22) above.
- Calibration can be estimated by a form of **cross-validation:** with a given prior I can (a) repeatedly divide the data at random into modeling and validation subsets, (b) update to posterior predictive distributions based on the modeling data, and (c) compare these distributions with the actual values in the validation data.

Note that calibration is **inherently frequentist** in spirit (e.g., "What percentage of the time do my 95% predictive intervals include the actual value?").

This leads to a useful **synthesis** of Bayesian and frequentist thinking:

> **Coherence** keeps me internally honest; **calibration** keeps me in good contact with the world.

## 2.6 Conjugate analysis; comparison with frequentist modeling

**Example:** 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).

To quantify these judgments I seek a **flexible family of densities** on $(0, 1)$,

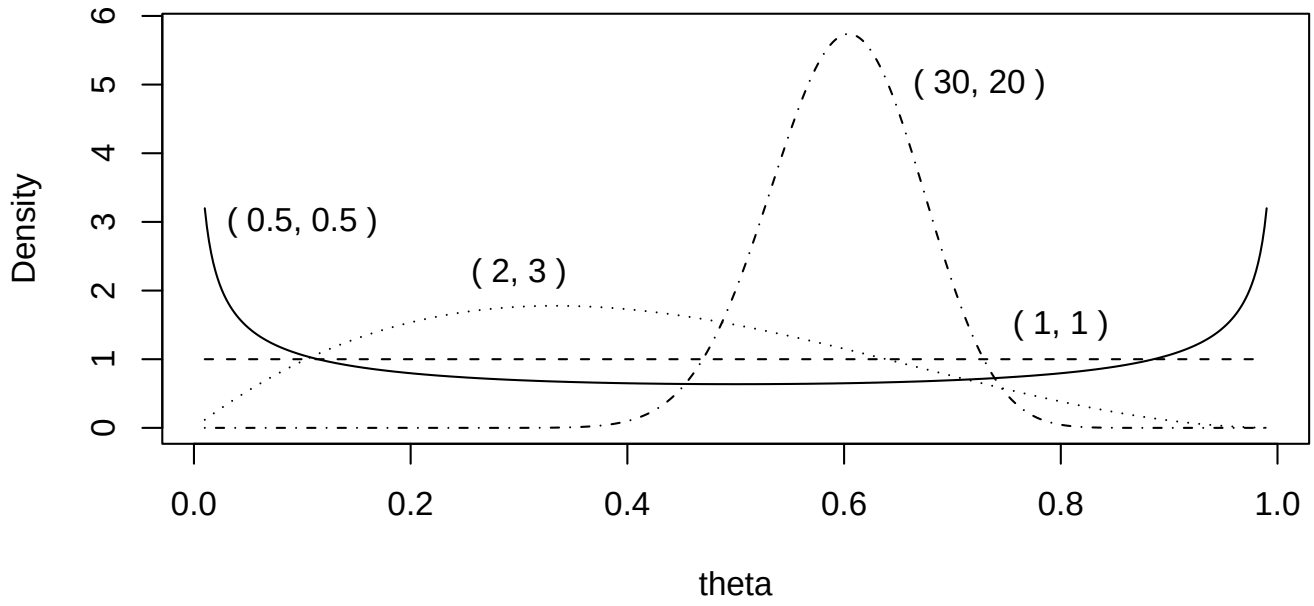
# The Beta Family of Densities on $(0, 1)$

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,

$$\text{Beta}(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\,\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{29}$$

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**:



# Conjugate Analysis (continued)

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

$$p(y_1, \ldots, y_n|\theta) = l(\theta|y) = \theta^s(1-\theta)^{n-s}, \tag{30}$$

where $s$ is the **sum** of the $y_i$.

Now Bayes' 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.

Rev. Bayes himself noticed back in the 1740s 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

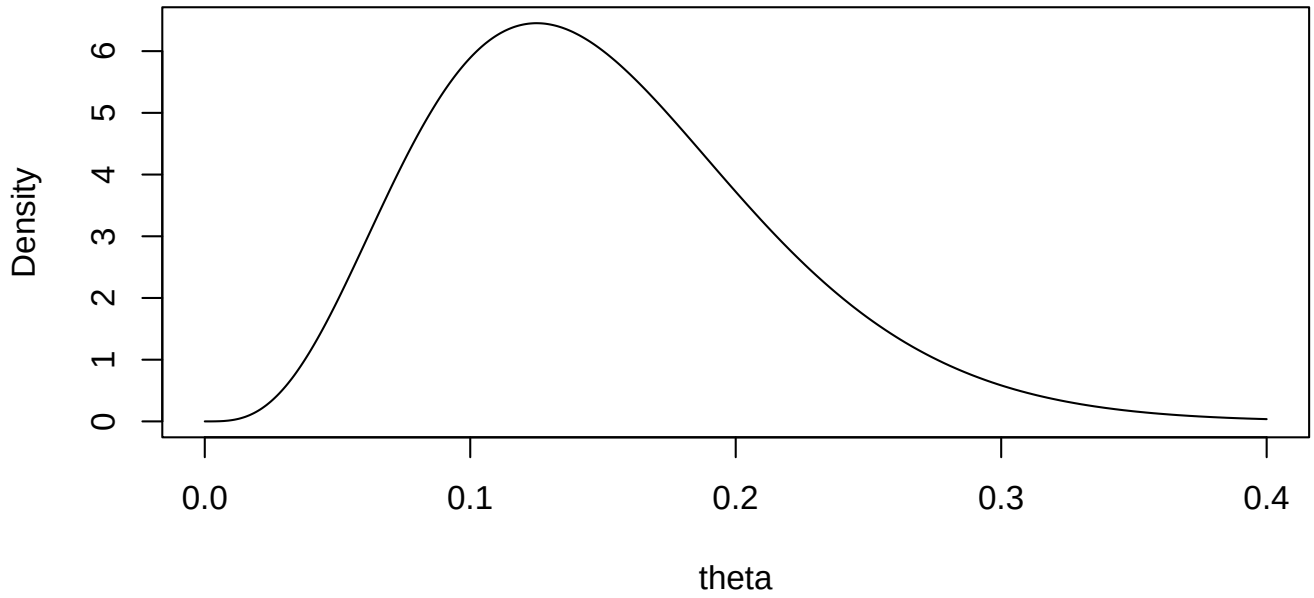
$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{31}$$

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

$$0.95 = \int_{0.05}^{0.30} \text{Beta}\left(\theta \,\middle|\, \alpha, \frac{17}{3}\alpha\right) d\theta \tag{32}$$

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

$$\begin{array}{rcll} (\alpha, \beta) & = & (4.5, 25.5) & \text{(hyperparameters)} \\ (\theta|\alpha, \beta) & \sim & \text{Beta}(\alpha, \beta) & \text{(prior)} \\ (Y_1, \ldots, Y_n|\theta) & \overset{\text{IID}}{\sim} & \text{Bernoulli}(\theta) & \text{(likelihood)} \end{array} \tag{33}$$

(33) 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 (I'll give more examples of this later).

**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 $\theta$, as noted above, with the Bernoulli likelihood),

$$\begin{aligned} p(\theta|y,\alpha,\beta) &= c\, l(\theta|y)\, p(\theta|\alpha,\beta) \\ &= c\, \theta^s\, (1-\theta)^{n-s}\, \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= c\, \theta^{(s+\alpha)-1} (1-\theta)^{(n-s+\beta)-1}, \end{aligned} \tag{34}$$

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 (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 $(4.5 + 25.5) = \mathbf{30}$ 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**.

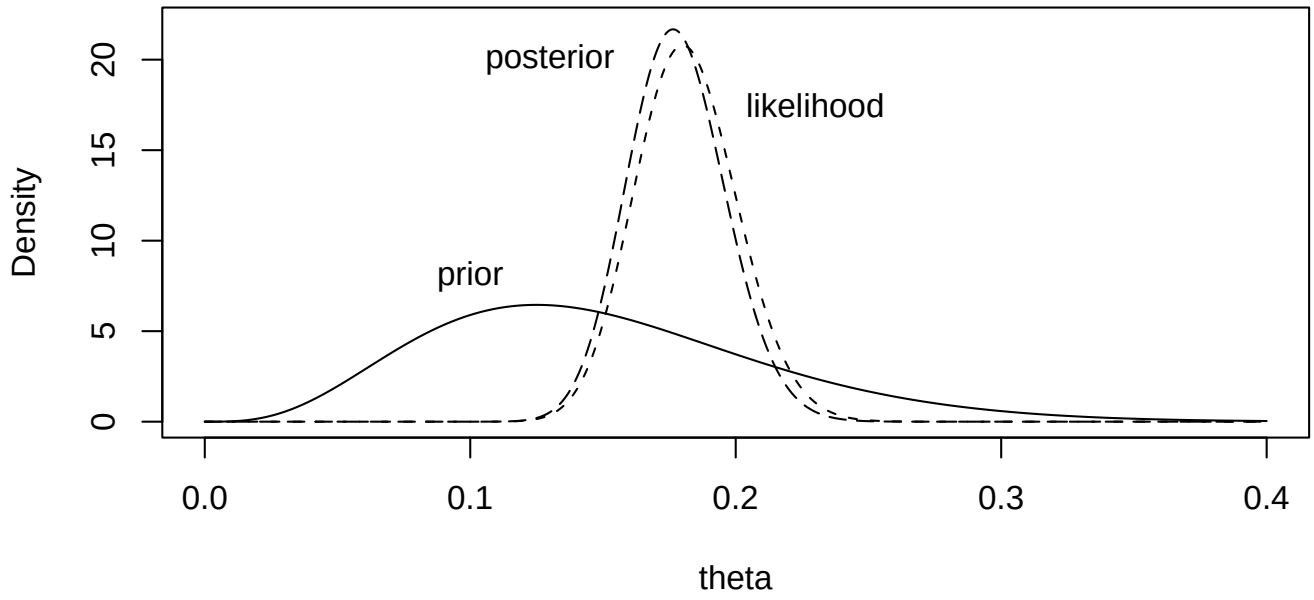
# Prior-To-Posterior Updating

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

$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\theta|y) \sim \text{Beta}(\alpha + s, \beta + n - s), \qquad (35)$$

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 $\theta$ 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** $n$:

$$\frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)\left(\frac{\alpha}{\alpha + \beta}\right) + \left(\frac{n}{\alpha + \beta + n}\right)\left(\frac{s}{n}\right)$$

$$\begin{array}{ccccccc} \text{posterior} & & \left(\begin{array}{c}\text{prior}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{prior}\\ \text{mean}\end{array}\right) & + & \left(\begin{array}{c}\text{data}\\ \text{weight}\end{array}\right) & \left(\begin{array}{c}\text{data}\\ \text{mean}\end{array}\right) \\ \text{mean} & = & & & & & \\ .178 & = & (.070) & (.15) & + & (.93) & (.18) \end{array}$$

# Comparison With Frequentist Modeling

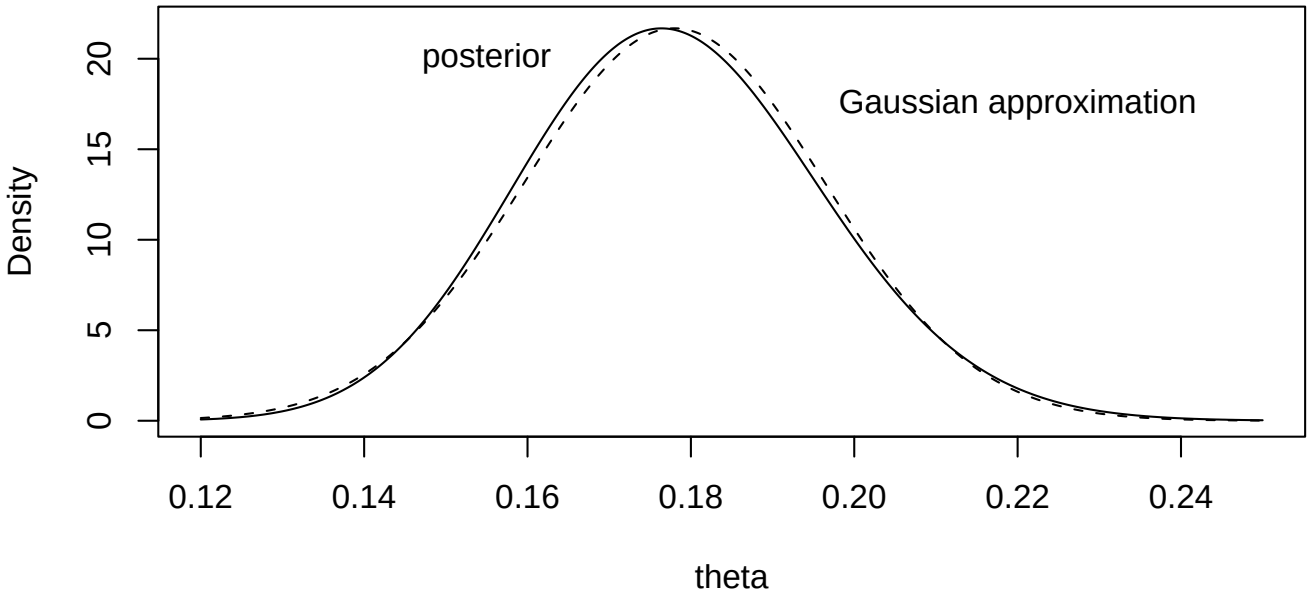
Another way to put this is that the data mean, $\bar{y} = \frac{s}{n} = \frac{72}{400} = .18$, has been **shrunk** toward the prior mean .15 by (in this case) a modest amount: the posterior mean is about .178, and the **shrinkage factor** is $\frac{30}{30+400}$ = about .07.

**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 $\mathrm{N}\left[\theta, \frac{\theta(1-\theta)}{n}\right]$, 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 $(\bar{Y} - \theta) \dot{\sim} \text{Gaussian}(0, .019^2)$; whereas

# Comparison With Frequentist Modeling (continued)

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

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**.

When **nondiffuse** prior information is available and I use it, my answer will **differ** from a frequentist analysis based on the same likelihood.

If my prior is retrospectively seen to have been **well-calibrated** I'll get a **better** answer than with the frequentist approach; if poorly calibrated, a **worse** answer (Samaniego and Reneau, 1994):

# Comparison With Frequentist Modeling (continued)

"bad" Bayesian $\leq$ frequentist $\leq$ "good" Bayesian

What you make of this depends on your **risk-aversion**: Is it better to try to land on the right in this box, running some risk of landing on the left, or to steer a middle course?

(**NB** I'll give several examples later in which a Bayesian analysis is better **even with diffuse prior information**: the point is that **likelihood methods don't always have good repeated-sampling properties with small samples**, and the **Bayesian approach** can **remedy** this problem.)

---

**Bernoulli prediction.** The **predictive distribution** for future $Y_i$ in the Bernoulli model was shown back on page 31 (equation (27)) to be

$$p(Y_{m+1} = y_{m+1}, \ldots, Y_n = y_n | y_1, \ldots, y_m) = \tag{36}$$
$$= \int_0^1 \prod_{i=m+1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \, p(\theta | y_1, \ldots, y_m) \, d\theta \ .$$

# Bernoulli Prediction (continued)

It became clear earlier that if the **prior** is taken to be Beta$(\alpha, \beta)$ the **posterior** $p(\theta|y_1, \ldots, y_m)$ in this expression is Beta$(\alpha^*, \beta^*)$, where $\alpha^* = \alpha + s$ and $\beta^* = \beta + (n - s)$.

As an example of an **explicit calculation** with (36) in this case, suppose that I've observed $n$ of the $Y_i$, obtaining data vector $y = (y_1, \ldots, y_n)$, and I want to predict $Y_{n+1}$.

Obviously $p(Y_{n+1} = y_{n+1}|y)$ has to be a **Bernoulli**$(\theta^*)$ distribution for some $\theta^*$, and intuition says that $\theta^*$ should just be the **mean** $\frac{\alpha^*}{\alpha^*+\beta^*}$ of the posterior distribution for $\theta$ given $y$.

(36) in this case gives for $p(Y_{n+1} = y_{n+1}|y)$ the expression

$$\int_0^1 \theta^{y_{n+1}} (1-\theta)^{1-y_{n+1}} \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \theta^{\alpha^*-1} (1-\theta)^{\beta^*-1}\, d\theta \tag{37}$$

$$= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\,\Gamma(\beta^*)} \int_0^1 \theta^{\alpha^*+y_{n+1}-1} (1-\theta)^{(\beta^*-y_{n+1}+1)-1}\, d\theta, \tag{38}$$

and a **symbolic computing package** such as `Maple` (or examination of the

# Bernoulli Prediction (continued)

logic leading to the **normalizing constant** of the **Beta distribution**) then yields that $p(Y_{n+1} = y_{n+1}|y)$ is

$$\left[\frac{\Gamma(\alpha^* + y_{n+1})}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\beta^* - y_{n+1} + 1)}{\Gamma(\beta^*)}\right]\left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right]. \tag{39}$$

Recalling that $\frac{\Gamma(x+1)}{\Gamma(x)} = x$ for any real number $x$ leads to **simple expressions** that **match intuition**; in the case $y_{n+1} = 1$, for instance, (39) becomes

$$\begin{aligned} p(Y_{n+1} = 1|y) &= \left[\frac{\Gamma(\alpha^* + 1)}{\Gamma(\alpha^*)}\right]\left[\frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^* + \beta^* + 1)}\right] \\ &= \frac{\alpha^*}{\alpha^* + \beta^*}. \end{aligned} \tag{40}$$

For example, with $(\alpha, \beta) = (4.5, 25.5)$ and $n = 400$ with $s = 72$, we saw earlier that the **posterior** for $\theta$ was Beta$(76.5, 353.5)$, and this posterior distribution has mean $\frac{\alpha^*}{\alpha^* + \beta^*} = 0.178$.

In this situation I would expect the next AMI patient who comes along to die within 30 days of admission with probability **0.178**, so the predictive distribution above **makes good sense**.

# The Binomial Distribution

It became clear above that a **sufficient statistic** for $\theta$ with a Bernoulli likelihood is the **sum** $s = \sum_{i=1}^{n} y_i$ of the 1s and 0s.

This means that if I buy into the model $(Y_i|\theta) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$, **I don't care** whether I observe the entire data vector $Y = (Y_1, \ldots, Y_n)$ or its sum $S = \sum_{i=1}^{n} Y_i$.

The distribution of $S$ in repeated sampling has a **familiar form**: it's just the **binomial** distribution $\text{Binomial}(n, \theta)$, which counts the number of successes in a series of IID success/failure trials.

Recall that if $S \sim \text{Binomial}(n, \theta)$ then $S$ has **discrete density**

$$p(S = s|\theta) = \left\{ \begin{array}{cc} \left( \begin{array}{c} n \\ s \end{array} \right) \theta^s (1-\theta)^{n-s} & \text{if } s = 0, \ldots, n \\ 0 & \text{otherwise} \end{array} \right\}.$$

This gives **another conjugate updating rule** in simple Bayesian modeling for free: if the data set just consists of a single draw $S$ from a binomial distribution, then the conjugate prior for the success probability $\theta$ is $\text{Beta}(\alpha, \beta)$, and the updating rule, which follows directly from (35), is

# Two Important General Points

$$\left\{ \begin{array}{c} \theta \sim \text{Beta}(\alpha, \beta) \\ (S|\theta) \sim \text{Binomial}(n, \theta) \end{array} \right\} \rightarrow (\theta|s) \sim \text{Beta}(\alpha + s, \beta + n - s). \qquad (41)$$

---

**1** (the **sequential** nature of **Bayesian learning**) Suppose you and I are observing data $(y_1, \ldots, y_n)$ to **learn** about a **parameter** $\theta$, and we have no reason throughout this observation process to **change** (the sampling distribution/likelihood part of) our **model**.

We both start with the **same prior** $p_1(\theta)$ before any of the data arrive, but we adopt what appear to be **different analytic strategies**:

- You wait until the whole data set $(y_1, \ldots, y_n)$ has been observed and **update** $p_1(\theta)$ **directly** to the posterior distribution $p(\theta|y_1, \ldots, y_n)$, whereas
- I **stop** after seeing $(y_1, \ldots, y_m)$ for some $m < n$, update $p_1(\theta)$ to an **intermediate** posterior distribution $p(\theta|y_1, \ldots, y_m)$, and then I go on from there, observing $(y_{m+1}, \ldots, y_n)$ and finally updating to a posterior on $\theta$ that takes account of the **whole data set** $(y_1, \ldots, y_n)$.

# Two Important General Points (continued)

**Q**$_1$ What should I use for my **intermediate prior distribution** $p_2(\theta)$?

**A**$_1$ Naturally enough, the **right thing to do** is to set $p_2(\theta) = p(\theta|y_1, \ldots, y_m)$.

The informal way people refer to this is to say that **yesterday's posterior distribution is today's prior distribution**.

**Q**$_2$ If I use the posterior in **A**$_1$, do you and I get the **same answer** for $p(\theta|y_1, \ldots, y_n)$ in the end?

**A**$_2$ **Yes** (you can check this).

**2** (the generality of **conjugate analysis**) Having seen **conjugate priors** used with binary outcomes, it's clear that **conjugate analysis** has a variety of **advantages**:

- It's **mathematically straightforward**;
- The **posterior mean** turns out to be a **weighted average** of the **prior** and **data means**; and
- The **prior** is nicely **interpretable** as an information source that's **equivalent to a data set**, and it's easy to figure out the **prior sample size**.

# Two Important General Points (continued)

It's natural to wonder, though, what's **lost** in addition to what's **gained** by adopting a conjugate prior.

The main **disadvantage** of conjugate priors is that in their simplest form they're **not flexible enough** to express **all possible forms** of prior information.

For example, in the AMI mortality case study, what if I wanted to combine a **bimodal** prior distribution with the Bernoulli likelihood?

This isn't possible when using a **single member** of the Beta$(\alpha, \beta)$ family.

However, it's possible to **prove** the following:

**Theorem** (Diaconis and Ylvisaker 1985). Given a particular likelihood that's a member of the **exponential family**, any prior distribution can be expressed as a **mixture** of priors that are conjugate to that likelihood.

For example, in the **AMI case study** the model could be

$$\begin{aligned} J &\sim p(J) \\ (\theta|J) &\sim \text{Beta}(\alpha_J, \beta_J) \end{aligned} \tag{42}$$

# Pros and Cons of Maximum Likelihood

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

for some distribution $p(J)$ on the positive integers — this is **completely general** but loses some of the advantages of simple conjugate analysis (e.g., **closed-form computations** are no longer possible).

---

**Pros and cons of maximum likelihood.** **Strength** of **maximum likelihood** as an approach to **parametric inference**:

- Fisher's approach (as you know) **extends** readily to situations in which the parameter $\theta$ is a **vector** of length $k > 1$ — the analogue of $\hat{I}$ in

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

  is a **matrix** (minus the **Hessian** [matrix of second partial derivatives] of the **log likelihood**, evaluated at the MLE) — so **maximum likelihood** is a **successful general approach** to **parametric inference** when the sample size $n$ is **large** and **little or no information external to the data set** $y$ is available (in this case **maximum likelihood** and **Bayesian inferential conclusions** will tend to be **similar**).

# Pros and Cons of Maximum Likelihood (continued)

**Disadvantages** of maximum likelihood in relation to **Bayesian inference** (as will become clear later):

- With **small samples sizes**, when the **likelihood function** $l(\theta|y)$ is **skewed** (e.g., often in **hierarchical models** [more on this later]), **maximization** over $\theta$ is **not the best technology** for learning about $\theta$; the **Bayesian** approach, which treats the likelihood as if it were a **density**, substitutes **integration** for **maximization** over $\theta$, and this has been found to have **better repeated-sampling properties** (with **diffuse priors**) when $n$ is small.
- The **frequentist** approach encourages thinking of each data set **in isolation**; the **Bayesian** approach explicitly provides a mechanism for **combining information** from **multiple sources**.
- **Prediction** of **observables** — an activity of **central importance** in **science/statistics** for its role in **model-checking** — is **much easier** from the **Bayesian** point of view.

# 2.7 References


Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.

Craig PS, Goldstein M, Seheult AH, Smith JA (1997). Constructing partial prior specifications for models of complex physical systems. *The Statistician*, **46**, forthcoming.

Draper D (1995). Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115–147, 233–239.

Draper D, Hodges JS, Mallows CL, Pregibon D (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**, 9–37.

de Finetti B (1930). Funzione caratteristica di un fenomeno aleatorio. *Mem. Acad. Naz. Lincei*, **4**, 86–133.

de Finetti B (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability*, HE Kyburg, Jr., and HE Smokler, eds., New York: Wiley (1980), 93–158.

de Finetti B (1974/5). *Theory of Probability*, **1–2**. New York: Wiley.

Fisher RA (1922). On the mathematical foundations of theoretical statistics.

# References (continued)

*Philosophical Transactions of the Royal Society of London A*, **222**, 309–368.

Fisher RA (1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.

Freedman D, Pisani R, Purves R, Adhikari A (1998). *Statistics*, third edition. New York: Norton.

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Bayesian Data Analysis*, second edition. London: Chapman & Hall.

Hacking I (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Johnson NL, Kotz S (1970). *Distributions in statistics: Continuous univariate distributions*, **1**. New York: Wiley.

Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**, 845–854.

Kadane JB, Wolfson LJ (1997). Experiences in elicitation. *The Statistician*, **46**, forthcoming.

Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990).

# References (continued)

The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953–1955 (with editorial comment, 1995–1997).

Laplace PS (1774). Mémoire sur la probabilité des causes par les évenements. *Mémoires de l'Academie des Sciences de Paris*, **6**, 621–656. English translation in 1986 as "Memoir on the probability of the causes of events," with an introduction by SM Stigler, *Statistical Science*, **1**, 359–378.

O'Hagan A (1997). Eliciting expert beliefs in substantial practical applications. *The Statistician*, **46**, forthcoming.

Samaniego FJ, Reneau DM (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*, **89**, 947–957.

Tierney L, Kadane JB (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.