over possible values of the unobserved data to obtain an observed data distribution

$$[\boldsymbol{Y}^{\text{obs}}|\boldsymbol{\theta}] = \int [\boldsymbol{Y}^{\text{obs}}, \boldsymbol{Y}^{\text{mis}}|\boldsymbol{\theta}]\, d\boldsymbol{Y}^{\text{mis}}. \tag{8.1}$$

The resulting likelihood $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{Y}^{\text{obs}}) = k\,[\boldsymbol{Y}^{\text{obs}}|\boldsymbol{\theta}]$ is the ODL.

An alternative approach is to treat the unobserved $\boldsymbol{Y}^{\text{mis}}$ in $[\boldsymbol{Y}^{\text{obs}}, \boldsymbol{Y}^{\text{mis}}|\boldsymbol{\theta}]$ exactly as we treat parameters $\boldsymbol{\theta}$. That is, we can treat $\boldsymbol{Y}^{\text{obs}}$ as fixed, and compare the support for various values of $\boldsymbol{Y}^{\text{mis}}$ and $\boldsymbol{\theta}$. Thus, we define a CDL as

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{Y}^{\text{mis}}|\boldsymbol{Y}^{\text{obs}}) \propto [\boldsymbol{Y}^{\text{obs}}, \boldsymbol{Y}^{\text{mis}}|\boldsymbol{\theta}].$$

There are a number of benefits to using CDL's rather than ODL's. First, the ODL can be a much more complicated function of parameters than the CDL. Second, use of the CDL provides a natural framework for prediction of unobserved quantities; describing missing data in terms of CDL's also tends to clarify assumptions required for predictions. CDL's are sometimes used in frequentist analyses (e.g., the EM algorithm, Dempster *et al.*, 1977). However, they are much more naturally and easily handled under the Bayesian paradigm, in which all unobserved quantities – be they parameters, predictions, missing values, whatever – all are treated equally (Section 5.1).

Before presenting examples, we take a moment to mention *data augmentation*, a concept closely related to that of CDL's, and one which users of Bayesian methods are likely to encounter. Data augmentation consists of expanding a model for observed data in terms of unobserved structures, usually with the goal of creating computational efficiencies. The augmented model includes a data distribution for observed data and the unobserved structures; integrating over values of the augmenting variables produces the ODL, as in Eq. (8.1). An example of data augmentation is found in Section 8.4.

8.2 RANDOMIZED RESPONSE DATA

One of our favorite demonstrations of the difference between ODLs and CDLs involves data from a randomized response survey. Randomized response surveys are used to estimate population rates of stigmatized behaviors, without requiring specific information about individuals surveyed. Such surveys were first described by Warner (1965), who begins thus:

> "For reasons of modesty, fear of being thought bigoted, or merely a reluctance to confide secrets to strangers, many individuals attempt to evade certain questions put to them by interviewers … either refusing outright to be surveyed, or consenting to be surveyed but purposely providing wrong answers … The questions that people tend to evade are the questions which demand answers that are too revealing."

Warner (1965) provided a clever solution. His technique has been applied in studies of alcohol use and abuse, smoking, abortion, academic cheating, criminal recidivism, sexual habits, and regulatory compliance (ranging from payment of taxes to obedience to fishing regulations). Brewer (1981) reports on its use by the Australian Bureau of Statistics, acting on behalf of the South Australian Royal Commission, in an investigation of marijuana use in Canberra.

Teaching a workshop for some of our colleagues — details of time and place suppressed, to protect the innocent — we wondered whether any of them might have ever engaged in Behavior X.[1] So we gathered some data.

We could not directly ask the question of interest because our colleagues may have been reluctant to own up to Behavior X. Instead, participants were asked to first toss a coin, without informing us of the outcome. They were asked to respond "Yes" if they had X'ed or if their coin came up heads. Those whose coin came up tails and who had not X'ed were asked to respond "No."

In this survey design, the data are deliberately confounded with the outcome of the coin toss in order to ensure that it is impossible to tell whether a person answering "Yes," responded positively because they had X'ed or whether it was because they obtained "Heads" on the coin flip. Despite this confounding, we can still make inference about the rate of Behavior X in the population and in our study group, modeling the confounding process using a binomial distribution.

It is easy to see how the probability of Behavior X can be estimated by considering a 2×2 table for the possible outcomes. Letting p denote the probability of Behavior X, and π denote the probability associated with the randomizing mechanism, and assuming the two events are independent, we calculate probabilities of four outcomes in Table 8.1.[2] In our randomized response data, three of the four outcomes (Heads and Behavior X, Tails and Behavior X, and Heads and no Behavior X) are all subsumed into the one response "Yes." The probability that a person answers "Yes" is $\theta = \pi + (1-\pi)p$; the probability of "No" is $(1-\pi)(1-p)$.

We can solve θ for p, obtaining $p = (\theta - \pi)/(1-\pi)$. Letting x denote the number responding "Yes" in a sample of n individuals, the maximum likelihood estimator (MLE) of θ is $\hat{\theta} = x/n$. Thus, given that $\hat{\theta} \geq \pi$, it follows from the invariance property of maximum likelihood

TABLE 8.1 Cell and margin probabilities for the 2×2 table of possible outcomes in the randomized-response to the Behavior X question.

	Behavior X		
	Yes	**No**	
Heads	πp	$\pi(1-p)$	π
Tails	$(1-\pi)p$	$(1-\pi)(1-p)$	$1-\pi$
	p	$(1-p)$	

Parameter π is the probability that the coin comes up Heads and p is the probability of Behavior X. Shaded cells correspond to response "Yes" in randomized-response survey, and have total probability $\theta = \pi + (1-\pi)p$.

1. We can't say what it was. Readers are asked to use their imagination.
2. We shall assume the coins flipped to be fair, hence $\pi = 1/2$. Randomization schemes could use alternative mechanisms, leading to alternative values of π; the important feature is that π is known.

estimation that the MLE of p is

$$\hat{p} = \frac{\hat{\theta} - \pi}{1 - \pi} = \frac{x - n\pi}{n(1 - \pi)}.$$

If $\hat{\theta} < \pi$, that is, if there are fewer positive responses than attributable to the randomizing event, the MLE of p is 0.[3]

From a modeling point of view it is instructive to think of how we would simulate the data. First, we would need to assign each subject a Behavior X status, using an indicator variable D_i for individual i. The next step would be to simulate the outcome of the individual's coin flip, with indicator variable C_i for the outcome "Heads." Because of the randomized response both D_i and C_i are latent variables, meaning they are potentially unobservable. What we always observe, however, is a variable Y_i which equals 0 if person i obtained tails and had never X'ed. Otherwise, Y_i is 1. We can write Y_i as a function of D_i and C_i, namely

$$Y_i = 1 - (1 - D_i)(1 - C_i). \tag{8.2}$$

If our n individuals can be regarded as a random sample from a large population, a reasonable way to simulate a value for D_i is to generate a Bernoulli random variable with success parameter p. Having simulated values for D_i, we next draw values C_i from a Bernoulli distribution with success parameter π. The observed variables Y_i are then constructed from D_i and C_i as in Eq. (8.2).

We represent this process using the *directed acyclic graph* (DAG) in Fig. 8.1. DAG's are an intuitive means for depicting relationships among quantities in hierarchical models. Single arrows denote stochastic dependencies (e.g., Bernoulli trials D have probability distributions depending on success parameters p) while double arrows denote deterministic relationships (e.g., Y is exactly calculated from D and C). The independence of D and C is indicated by the absence of shared stochastic dependencies. The DAG is *directed*: we write $p \to D$ because our model will describe the relation between p and D in terms of $[D|p]$ rather than the other way around. The DAG is *acyclic* in that we avoid model specifications of the form $X \to Y \to Z \to X$ which cycle back on themselves, and might wind up describing models that do not make sense.[4]

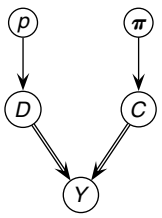

FIGURE 8.1 Directed acyclic graph (DAG) representation of the complete data likelihood for the Behavior X example.

3. The MLE is biased because of the truncation. Warner (1965) describes the estimator as unbiased, but the result requires the acceptance of negative estimates.
4. For example, $[X|Y] = N(Y, 1)$ and $[Y|X] = N(X, 1)$ might seem a reasonable model specification, but there is no joint probability distribution on $\{X, Y\}$, which yields these conditional distributions.

In this particular problem, the variables C and D are partially observed through Y. For cases where $Y_i = 0$, we know that $C_i = D_i = 0$. If we use the superscript obs to denote observed values and mis to denote missing values, we can write the CDL as:

$$\begin{aligned}\mathcal{L}(p, \boldsymbol{C}^{\text{mis}}, \boldsymbol{D}^{\text{mis}} | \boldsymbol{Y}) &\propto [\boldsymbol{Y}, \boldsymbol{C}, \boldsymbol{D} | p] \\ &= \prod_{i=1}^{n} I(Y_i, C_i, D_i) \times [\boldsymbol{C}]\,[\boldsymbol{D} | p] \\ &= \prod_{i=1}^{n} I(Y_i, C_i, D_i) \times \pi^{C_i}(1-\pi)^{1-C_i} \times p^{D_i}(1-p)^{1-D_i}.\end{aligned}$$

Here, $I(Y_i, C_i, D_i)$ is an indicator for the defining relationship given in Eq. (8.2). Thus, $I(Y_i, C_i, D_i) = 1$ if either (1) $Y_i = 1$ and $\{C_i, D_i\} \neq \{0, 0\}$ or (2) $Y_i = 0$, $C_i = 0$, and $D_i = 0$. Otherwise $I(Y_i, C_i, D_i) = 0$. The role of this indicator function is to enforce constraints that the data impose on allowable values for C_i and D_i, and in particular that we cannot have $Y_i = 1$, $C_i = 0$, and $D_i = 0$. This constraint defines the role of the observed data ($\boldsymbol{Y}$) in providing information about p in the CDL.

As a problem in Bayesian inference, we require the posterior distributions of all unknown quantities. This is proportional to the CDL multiplied by the joint prior on unknown parameters. If we also use a Be(α, β) prior for p, then

$$\left[p, \boldsymbol{C}^{\text{mis}}, \boldsymbol{D}^{\text{mis}} | \boldsymbol{Y}\right] \propto \left\{ \prod_{i=1}^{n} \left\{ I(Y_i, C_i, D_i) \right\} \left\{ \pi^{C_i}(1-\pi)^{1-C_i} \right\} \left\{ p^{D_i}(1-p)^{1-D_i} \right\} \right\} \times p^{\alpha-1}(1-p)^{\beta-1}. \tag{8.3}$$

8.2.1 Calculating Posterior Distributions

For inference about p, we can proceed in a number of ways. One approach is to try to find an explicit expression for the density $[p | \boldsymbol{Y}]$. Formally, we integrate (in this case sum) over the possible values for the latent variables C_i and D_i in Eq. (8.3). We spare the reader the ugly details, noting only that the result is the same as obtained by first computing the ODL based on Table 8.1, then multiplying by the prior for p.

We will follow that easier course. The values Y_i are exchangeable (conditionally independent) Bernoulli trials. Examining Table 8.1, we see that $\theta \equiv \Pr(Y_i = 1) = \pi + (1-\pi)p$; thus, $X = \sum_{i=1}^{n} Y_i$ is a binomial random variable with index n and success rate θ. The ODL is thus $\mathcal{L}(p | \boldsymbol{Y}) \propto \theta^X (1-\theta)^{(n-X)}$. Multiplying by the prior for p, we have

$$[p | \boldsymbol{Y}] \propto (\pi + (1-\pi)p)^X ((1-\pi)(1-p))^{n-X} \times p^{\alpha-1}(1-p)^{\beta-1}. \tag{8.4}$$

The posterior density for p does not exist in closed-form for arbitrary α and β. While it is possible to find the required normalizing constant using software like R, an easier solution is desirable. Nevertheless, we have done so for two priors, graphing the densities in Fig. 8.2.

Analysis of the Behavior X data is made much easier through Markov chain Monte Carlo (MCMC). We present three different approaches, based on the CDL, the ODL, and a partially integrated version of the CDL.