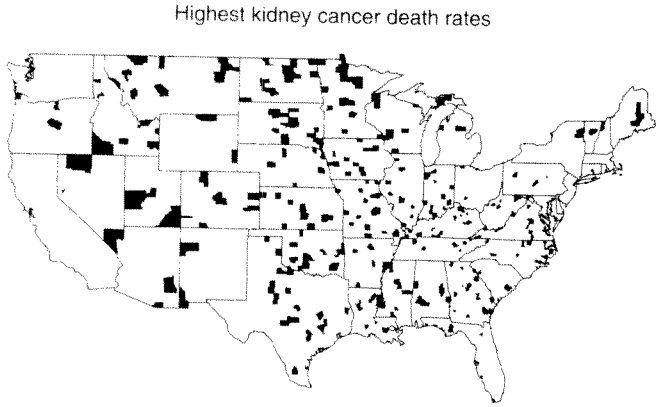

Figure 2.7 *The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Why are most of the shaded counties in the middle of the country? See Section 2.8 for discussion.*

each based on different data but with a common prior distribution. In addition to illustrating the role of the prior distribution, this example introduces hierarchical modeling, to which we return in Chapter 5.

A puzzling pattern in a map

Figure 2.7 shows the counties in the United States with the highest kidney cancer death rates during the 1980s.* The most noticeable pattern in the map is that many of the counties in the Great Plains in the middle of the country, but relatively few counties near the coasts, are shaded.

When shown the map, people come up with many theories to explain the disproportionate shading in the Great Plains: perhaps the air or the water is polluted, or the people tend not to seek medical care so the cancers get detected too late to treat, or perhaps their diet is unhealthy ... These conjectures may all be true but they are not actually needed to explain the patterns in Figure 2.7. To see this, look at Figure 2.8, which plots the 10% of counties with the *lowest* kidney cancer death rates. These are also mostly in the middle of the country. So now we need to explain why these areas have the lowest, as well as the highest, rates.

The issue is sample size. Consider a county of population 1000. Kidney cancer is a rare disease, and, in any ten-year period, a county of 1000 will probably have zero kidney cancer deaths, so that it will be tied for the lowest rate in the country and will be shaded in Figure 2.8. However, there is a chance

* The rates are age-adjusted and restricted to white males, issues which need not concern us here.

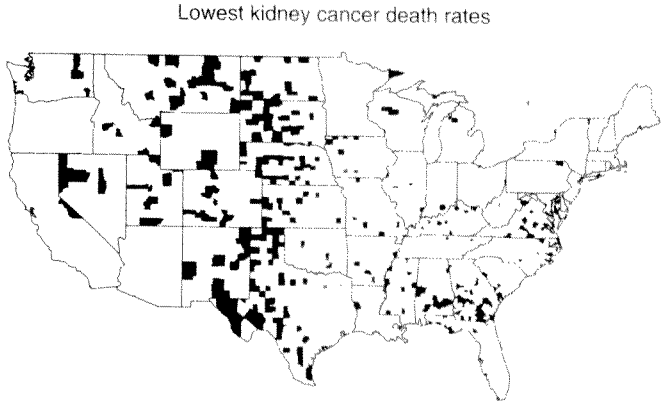

Figure 2.8 *The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Surprisingly, the pattern is somewhat similar to the map of the highest rates, shown in Figure 2.7.*

the county will have one kidney cancer death during the decade. If so, it will have a rate of 1 per 10,000 per year, which is high enough to put it in the top 10% so that it will be shaded in Figure 2.7. The Great Plains has many low-population counties, and so it is disproportionately represented in both maps. There is no evidence from these maps that cancer rates are particularly high there.

Bayesian inference for the cancer death rates

The misleading patterns in the maps of raw rates suggest that a model-based approach to estimating the true underlying rates might be helpful. In particular, it is natural to estimate the underlying cancer death rate in each county j using the model

$$y_j \sim \text{Poisson}(10 n_j \theta_j), \tag{2.16}$$

where y_j is the number of kidney cancer deaths in county j from 1980–1989, n_j is the population of the county, and θ_j is the underlying rate in units of deaths per person per year. (Here we are ignoring the age-standardization, although a generalization of the model to allow for this would be possible.)

This model differs from (2.14) in that θ_j varies between counties, so that (2.16) is a separate model for each of the counties in the U.S. We use the subscript j (rather than i) in (2.16) to emphasize that these are separate parameters, each being estimated from its own data. Were we performing inference for just one of the counties, we would simply write $y \sim \text{Poisson}(10n\theta)$.

To perform Bayesian inference, we need a prior distribution for the unknown rate θ_j. For convenience we use a gamma distribution, which is conjugate to

the Poisson. As we shall discuss later, a gamma distribution with parameters $\alpha = 20$ and $\beta = 430{,}000$ is a reasonable prior distribution for underlying kidney cancer death rates in the counties of the U.S. during this period. This prior distribution has a mean of $\alpha/\beta = 4.65 \times 10^{-5}$ and standard deviation $\sqrt{\alpha}/\beta = 1.04 \times 10^{-5}$.

The posterior distribution of θ_j is then,

$$\theta_j | y_j \sim \text{Gamma}(20 + y_j,\ 430000 + 10n_j),$$

which has mean and variance,

$$\begin{aligned} \text{E}(\theta_j | y_j) &= \frac{20 + y_j}{430{,}000 + 10n_j} \\ \text{var}(\theta_j | y_j) &= \frac{20 + y_j}{(430{,}000 + 10n_j)^2}. \end{aligned}$$

The posterior mean can be viewed as a weighted average of the raw rate, $y_j/(10n_j)$, and the prior mean, $\alpha/\beta = 4.65 \times 10^{-5}$. (For a similar calculation, see Exercise 2.5.)

Relative importance of the local data and the prior distribution

Inference for a small county. The relative weighting of prior information and data depends on the population size n_j. For example, consider a small county with $n_j = 1000$:

- For this county, if $y_j = 0$, then the raw death rate is 0 but the posterior mean is $20/440{,}000 = 4.55 \times 10^{-5}$.
- If $y_j = 1$, then the raw death rate is 1 per 1000 per 10 years, or 10^{-4} per person-year (about twice as high as the national mean), but the posterior mean is only $21/440{,}000 = 4.77 \times 10^{-5}$.
- If $y_j = 2$, then the raw death rate is an extremely high 2×10^{-4} per person-year, but the posterior mean is still only $22/440{,}000 = 5.00 \times 10^{-5}$.

With such a small population size, the data are dominated by the prior distribution.

But how likely, *a priori*, is it that y_j will equal 0, 1, 2, and so forth, for this county with $n_j = 1000$? This is determined by the predictive distribution, the marginal distribution of y_j, averaging over the prior distribution of θ_j. As discussed in Section 2.7, the Poisson model with gamma prior distribution has a negative binomial predictive distribution:

$$y_j \sim \text{Neg-bin}\left(\alpha, \frac{\beta}{10n_j}\right).$$

It is perhaps even simpler to simulate directly the predictive distribution of y_j as follows: (1) draw 500 (say) values of θ_j from the Gamma(20, 430000) distribution; (2) for each of these, draw one value y_j from the Poisson distribution with parameter $10000\theta_j$. Of 500 simulations of y_j produced in this way, 319 were 0's, 141 were 1's, 33 were 2's, and 5 were 3's.

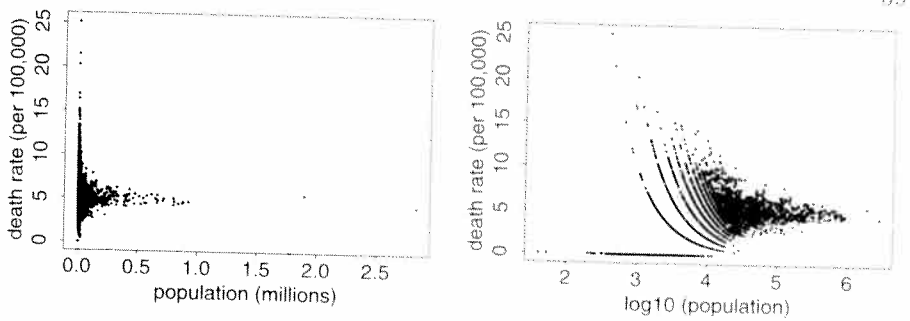

Figure 2.9 *(a) Kidney cancer death rates $y_j/(10n_j)$ vs. population size n_j. (b) Replotted on the scale of* $\log_{10}$ *population to see the data more clearly. The patterns come from the discreteness of the data ($n_j = 0, 1, 2, \ldots$).*

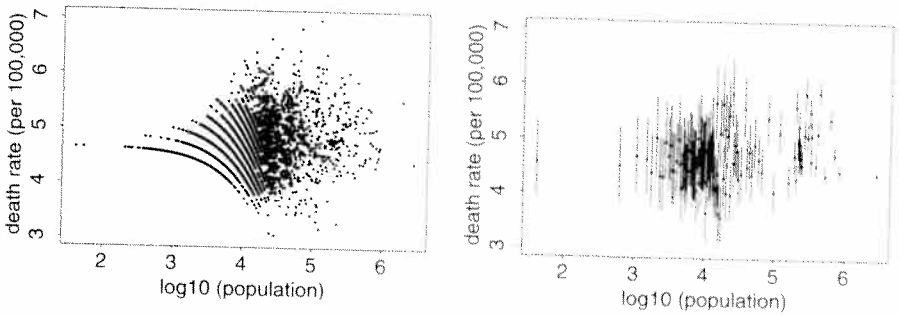

Figure 2.10 *(a) Bayes-estimated posterior mean kidney cancer death rates, $E(\theta_j|y_j) = \frac{20+y_j}{430{,}000+10n_j}$ vs. logarithm of population size n_j, the 3071 counties in the U.S. (b) Posterior medians and 50% intervals for θ_j for a sample of 100 counties j. The scales on the y-axes differ from the plots in Figure 2.9b.*

Inference for a large county. Now consider a large county with $n_j = 1$ million. How many cancer deaths y_j might we expect to see in a ten-year period? Again we can use the Gamma(20, 430000) and Poisson($10^7 \cdot \theta_j$) distributions to simulate 500 values of y_j from the predictive distribution. Doing this we found a median of 473 and a 50% interval of [393, 545]. The raw death rate is then as likely or not to fall between 3.93×10^{-5} and 5.45×10^{-5}.

What about the Bayesianly estimated or 'Bayes-adjusted' death rate? For example, if y_j takes on the low value of 393, then the raw death rate is 3.93×10^{-5} and the posterior mean of θ_j is $(20 + 393)/(10{,}430{,}000) = 3.96 \times 10^{-5}$, and if $y_j = 545$, then the raw rate is 5.45×10^{-5} and the posterior mean is 5.41×10^{-5}. In this large county, the data dominate the prior distribution.

Comparing counties of different sizes. In the Poisson model (2.16), the variance of y_j is inversely proportional to the exposure parameter n_j, which can thus be considered a 'sample size' for county j. Figure 2.9 shows how the raw kidney cancer death rates vary by population. The extremely high and

extremely low rates are all in low-population counties. By comparison, Figure 2.10a shows that the Bayes-estimated rates are much less variable. Finally, Figure 2.10b displays 50% interval estimates for a sample of counties (chosen because it would be hard to display all 3071 in a single plot). The smaller counties supply less information and thus have wider posterior intervals.

Constructing a prior distribution

We now step back and discuss where we got the Gamma(20.430000) prior distribution for the underlying rates. As we discussed when introducing the model, we picked the gamma distribution for mathematical convenience. We now explain how the two parameters α, β can be estimated from data to match the distribution of the observed cancer death rates $y_j/(10n_j)$. It might seem inappropriate to use the data to set the prior distribution, but we view this as a useful approximation to our preferred approach of hierarchical modeling (introduced in Chapter 5), in which distributional parameters such as α, β in this example are treated as unknowns to be estimated.

Under the model, the observed count y_j for any county j comes from the predictive distribution, $p(y_j) = \int p(y_j|\theta_j)p(\theta_j)d\theta_j$, which in this case is Neg-bin$(\alpha, \beta/(10n_j))$. From Appendix A, we can find the mean and variance of this distribution:

$$\begin{aligned} \mathrm{E}(y_j) &= 10n_j\frac{\alpha}{\beta} \\ \mathrm{var}(y_j) &= 10n_j\frac{\alpha}{\beta} + (10n_j)^2\frac{\alpha}{\beta^2}. \end{aligned} \tag{2.17}$$

These can also be derived directly using the mean and variance formulas (1.7) and (1.8); see Exercise 2.6.

Matching the observed mean and variance to their expectations and solving for α and β yields the parameters of the prior distribution. The actual computation is more complicated because it is better to deal with the mean and variance of the rates $y_j/(10n_j)$ than the y_j and we must also deal with the age adjustment, but the basic idea of matching moments presented here illustrates that the information is present to estimate α and β from the data of the 3071 counties.

Figure 2.11 shows the empirical distribution of the raw cancer rates, along with the estimated Gamma(20.430000) prior distribution for the underlying cancer rates θ_j. The distribution of the raw rates is much broader, which makes sense since they include the Poisson variability as well as the variation between counties.

Our prior distribution is reasonable in this example, but this method of constructing it – by matching moments– is somewhat sloppy and can be difficult to apply in general. In Chapter 5, we discuss how to estimate this and other prior distributions in a more direct Bayesian manner, in the context of hierarchical models.

A more important way this model could be improved is by including information at the county level that could predict variation in the cancer rates. This would move the model toward a hierarchical Poisson regression of the sort discussed in Chapter 16.

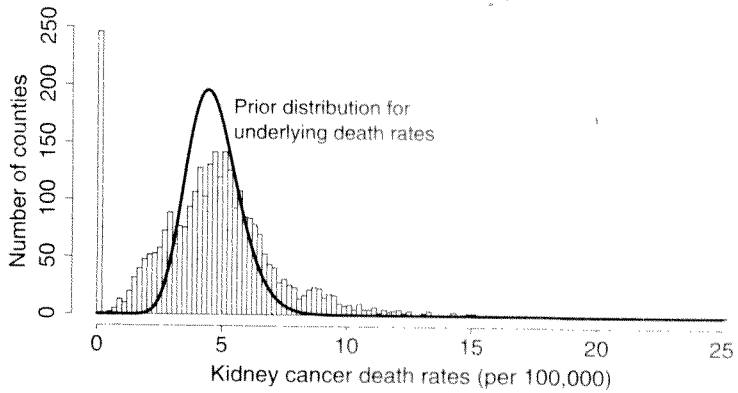

Figure 2.11 *Empirical distribution of the age-adjusted kidney cancer death rates, $y_j/(10n_j)$, for the 3071 counties in the U.S., along with the Gamma(20.430000) prior distribution for the underlying cancer rates θ_j.*

2.9 Noninformative prior distributions

When prior distributions have no population basis, they can be difficult to construct, and there has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution. Such distributions are sometimes called 'reference prior distributions,' and the prior density is described as vague, flat, diffuse or *noninformative*. The rationale for using noninformative prior distributions is often said to be 'to let the data speak for themselves,' so that inferences are unaffected by information external to the current data.

Proper and improper prior distributions

We return to the problem of estimating the mean θ of a normal model with known variance σ^2, with a $\mathrm{N}(\mu_0, \tau_0^2)$ prior distribution on θ. If the prior precision, $1/\tau_0^2$, is small relative to the data precision, n/σ^2, then the posterior distribution is approximately as if $\tau_0^2 = \infty$:

$$p(\theta|y) \approx \mathrm{N}(\theta|\overline{y}, \sigma^2/n).$$

Putting this another way, the posterior distribution is approximately that which would result from assuming $p(\theta)$ is proportional to a constant for $\theta \in (-\infty, \infty)$. Such a distribution is not strictly possible, since the integral of the assumed $p(\theta)$ is infinity, which violates the assumption that probabilities sum to 1. In general, we call a prior density $p(\theta)$ *proper* if it does not depend on data and integrates to 1. (If $p(\theta)$ integrates to any positive finite value, it