2. Single-parameter models

2020-24837 Kyungmin Lee

2020-30213 Jinwon Park

Contents

2.1 Estimating a probability from binomial data

Binomial sampling model - Likelihood, Posterior

$$p(y|\theta) = Bin(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \tag{2.1}$$

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y} \tag{2.2}$$

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

(Bayes' rule)

2.1 Estimating a probability from binomial data

Beta & Gamma distribution

$$Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad \theta \in [0,1]$$

$$Gamma(\alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\theta^{-(\alpha-1)}e^{-\beta\theta}, \qquad \theta > 0$$

2.1 Estimating a probability from binomial data

Binomial sampling model - Likelihood, Posterior

$$p(y|\theta) = Bin(y|n,\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y} \tag{2.1}$$

$$p(\theta|y) \propto \theta^{y}(1-\theta)^{n-y} \tag{2.2}$$

$$\theta\ |y \sim Beta(y+1, n-y+1) \tag{2.3}$$

2.1 Estimating a probability from binomial data

Example. Estimating the probability of a female birth

$$\theta \mid y \sim Beta(y+1, n-y+1)$$

n : population

y : # of female

θ : proportion of female births

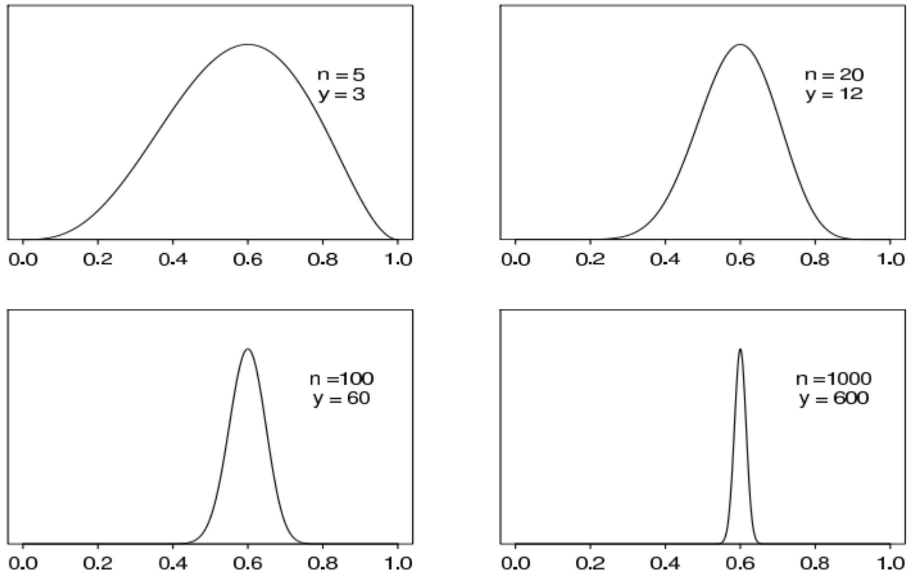


Figure 2.1 *Unnormalized posterior density for binomial parameter θ, based on uniform prior distribution and y successes out of n trials. Curves displayed for several values of n and y.*

Assume. $\theta \sim U(0,1)$

2.1 Estimating a probability from binomial data

Prediction

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y},\theta|y)d\theta \\
&= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta \\
&= \int p(\tilde{y}|\theta)p(\theta|y)d\theta
\end{aligned}
\tag{1.4}
$$

2.1 Estimating a probability from binomial data

Prediction

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

$$\begin{aligned} Pr(\tilde{y} = 1|y) &= \int_0^1 Pr(\tilde{y} = 1|\theta, y)p(\theta|y)d\theta \\ &= \int_0^1 \theta p(\theta|y)d\theta = E(\theta|y) = \frac{\alpha}{\alpha + \beta} = \frac{y+1}{n+2} \end{aligned} \tag{1.4}$$

2.2 Posterior as compromise between data and prior information

Relationship between prior & posterior mean & variance

$$E(\theta) = E\big(E(\theta|y)\big) \tag{2.7}$$

$$var(\theta) = E\big(var(\theta|y)\big) + var\big(E(\theta|y)\big) \tag{2.8}$$

$var(\theta) > E\big(var(\theta|y)\big)$: potential for reducing 'uncertainty'

2.3 Summarizing posterior inference

- Flexibility : posterior inferences can be summarized, even after complicated transformations

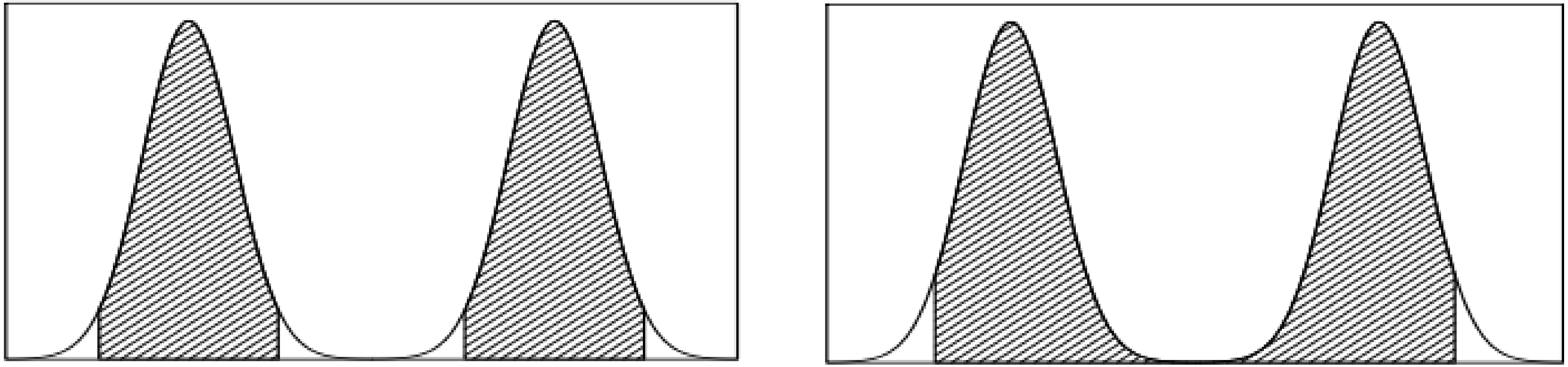
- Summaries of locations : mean, median, mode

- The mode often plays an important role even rather than mean or median because of its convenience

- Posterior quantiles : interest in interval summary with regard to posterior uncertainty

2.3 Summarizing posterior inference

- The highest posterior density region
  : conveys more information about separate centrals



(a) central posterior interval, (b) highest posterior density region.

2.4 Informative prior distributions

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

The uniform prior distribution

$$\theta \sim U(0,1) \quad \Rightarrow \quad p(\theta) = 1, \qquad p(\theta|y) \propto p(y|\theta)$$

Informative prior cases

$$\theta \sim \blacksquare \quad \Rightarrow \quad p(\theta) = \blacksquare, \qquad p(\theta|y) \propto \blacksquare \cdot p(y|\theta)$$

2.4 Informative prior distributions

Conjugacy : Binomial sampling model
with hyperparameter α, β of Beta distribution

Likelihood

$$p(y|\theta) \propto \theta^y (1-\theta)^{n-y}$$

Prior informative

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Posterior

$$\begin{aligned} p(\theta|y) &\propto \theta^y (1-\theta)^{n-y} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= Beta(\theta|\alpha+y, \beta+n-y) \quad \triangle \end{aligned}$$

2.4 Informative prior distributions

Definition of conjugacy

$$p(\theta|y) \in \mathcal{P} \text{ for all } p(\cdot\,|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

$\mathcal{F}$: class of $p(\theta|y)$, $\mathcal{P}$: class of all distribution

- Interested in natural conjugate prior families

2.4 Informative prior distributions

Exponential families

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}$$

$$p(y|\theta) \propto g(\theta)^n e^{\phi(\theta)^T t(y)}, \qquad \text{where } t(y) = \sum_{i=1}^{n} u(y_i)$$

- t(y) : sufficient statistic for θ
- The only classes that have natural conjugate prior

2.5 Estimating a normal mean with unknown variance

Normal sample distribution (known σ^2)

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$$

$\theta \sim N(\mu_0, {\tau_0}^2)$ with hyperparameters μ_0, ${\tau_0}^2$

$$p(\theta) \propto \exp(-\frac{1}{2{\tau_0}^2}(\theta - \mu_0)^2)$$

2.5 Estimating a normal mean with unknown variance

$\theta|y \sim N(\mu_1, {\tau_1}^2)$

$$p(\theta|y) \propto \exp(-\frac{1}{2{\tau_1}^2}(\theta - \mu_1)^2)$$

$$\mu_1 = \frac{\frac{1}{{\tau_0}^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{{\tau_0}^2} + \frac{1}{\sigma^2}} \text{ and } \frac{1}{{\tau_1}^2} = \frac{1}{{\tau_0}^2} + \frac{1}{\sigma^2} \text{ (precision)}$$

2.5 Estimating a normal mean with unknown variance

Compromise between the prior mean and the observed value

$$\mu_1 = \mu_0 + (y - \mu_0)\frac{{\tau_0}^2}{\sigma^2 + {\tau_0}^2}$$

$$\mu_1 = y - (y - \mu_0)\frac{\sigma^2}{\sigma^2 + {\tau_0}^2}$$

2.5 Estimating a normal mean with unknown variance

Posterior predictive distribution

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

$$\propto \int \exp(-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2)\exp(-\frac{1}{2{\tau_1}^2}(\theta - \mu_1)^2)d\theta$$

2.5 Estimating a normal mean with unknown variance

Posterior predictive distribution

$$E(\tilde{y}|y) = E(E(\tilde{y}|\theta, y)|y = E(\theta|y) = \ \mu_1$$

$$\begin{aligned} var(\tilde{y}|y) &= E(var(\tilde{y}|\theta, y)|y) + var(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + var(\theta|y) \\ &= \sigma^2 + {\tau_1}^2 \end{aligned}$$

2.5 Estimating a normal mean with unknown variance

Normal model with multiple observable

$$p(\theta|y_1 \dots, y_n) = p\left(\theta\middle|\bar{y}\right) = N(\theta|\mu_n,{\tau_n}^2)$$

$$\mu_n = \frac{\frac{1}{{\tau_0}^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{{\tau_0}^2} + \frac{n}{\sigma^2}} \text{ and } \frac{1}{{\tau_1}^2} = \frac{1}{{\tau_0}^2} + \frac{n}{\sigma^2} \text{ (precision)}$$

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

2.6 Other standard single-parameter models

Normal distribution with known mean but unknown variance

$$p(y|\sigma^2) \quad \propto \quad \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)$$

2.6 Other standard single-parameter models

Normal distribution with known mean but unknown variance

Likelihood

$$\begin{aligned} p(y|\sigma^2) &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2\sigma^2}v\right). \end{aligned}$$

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2.$$

v : sufficient static

2.6 Other standard single-parameter models

Normal distribution with known mean but unknown variance

Conjugate prior (Inverse-Gamma)

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

2.6 Other standard single-parameter models

Normal distribution with known mean but unknown variance

Conjugate posterior (Inverse-chi-square)

$$
\begin{aligned}
p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\
&\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\nu_0/2+1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right) \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}\frac{v}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-((n+\nu_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + nv)\right).
\end{aligned}
$$

2.6 Other standard single-parameter models

Normal distribution with known mean but unknown variance

Conjugate posterior (Inverse-chi-square)

$$\sigma^2|y \sim \text{Inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + nv}{\nu_0 + n}\right)$$

2.6 Other standard single-parameter models

Poisson model

Likelihood (y ~ Poisson(θ))

$$p(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}, \quad \text{for } y = 0, 1, 2, \ldots$$

2.6 Other standard single-parameter models

Poisson model

Likelihood : exponential family form

$$
\begin{aligned}
p(y|\theta) &= \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i} e^{-\theta} \\
&\propto \theta^{t(y)} e^{-n\theta}, \\
&\propto e^{-n\theta} e^{t(y)\log\theta}
\end{aligned}
$$

natural parameter

$$\phi(\theta) = \log\theta$$

2.6 Other standard single-parameter models

Poisson model

Prior predictive dist. – the negative binomial density

$$
\begin{aligned}
p(y) &= \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,1+\beta)} \\
&= \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}}, \\
&= \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y},
\end{aligned}
$$

$$y \sim \text{Neg-bin}(\alpha, \beta)$$

2.6 Other standard single-parameter models

Poisson model

Conjugate prior distribution

$$p(\theta) \propto (e^{-\theta})^{\eta} e^{\nu \log \theta}$$

Conjugate posterior distribution

$$\theta | y \sim \text{Gamma}(\alpha + n\overline{y}, \beta + n)$$

2.6 Other standard single-parameter models

Poisson model parameterized in terms of rate and exposure

Extension of Poisson model for data points $y_1, \ldots, y_n$

$$y_i \sim \text{Poisson}(x_i\theta)$$

x_i : exposure θ : rate

Likelihood

$$p(y|\theta) \propto \theta^{\left(\sum_{i=1}^{n} y_i\right)} e^{-\left(\sum_{i=1}^{n} x_i\right)\theta}$$

2.6 Other standard single-parameter models

Poisson model parameterized in terms of rate and exposure

Prior

$$\theta \sim \text{Gamma}(\alpha, \beta),$$

Posterior

$$\theta | y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^{n} y_i,\ \beta + \sum_{i=1}^{n} x_i\right)$$

2.6 Other standard single-parameter models

Exponential model – time scale, ‘waiting times’

Rate

$$\theta = 1/\mathrm{E}(y|\theta)$$

A sampling distribution (not used as a likelihood)

$$p(y|\theta) = \theta \exp(-y\theta), \text{ for } y > 0.$$

Prior and Posterior

$$Gamma(\theta|\alpha, \beta) \qquad Gamma(\theta|\alpha + 1, \beta + y)$$

2.6 Other standard single-parameter models

Exponential model – time scale, ‘waiting times’

Rate

$$\theta = 1/\mathrm{E}(y|\theta)$$

A sampling distribution (not used as a likelihood)

$$p(y|\theta) = \theta \exp(-y\theta), \text{ for } y > 0$$

Prior and Posterior

$$Gamma(\theta|\alpha, \beta) \qquad Gamma(\theta|\alpha + 1, \beta + y)$$

2.6 Other standard single-parameter models

Exponential model – n independent exp. observations

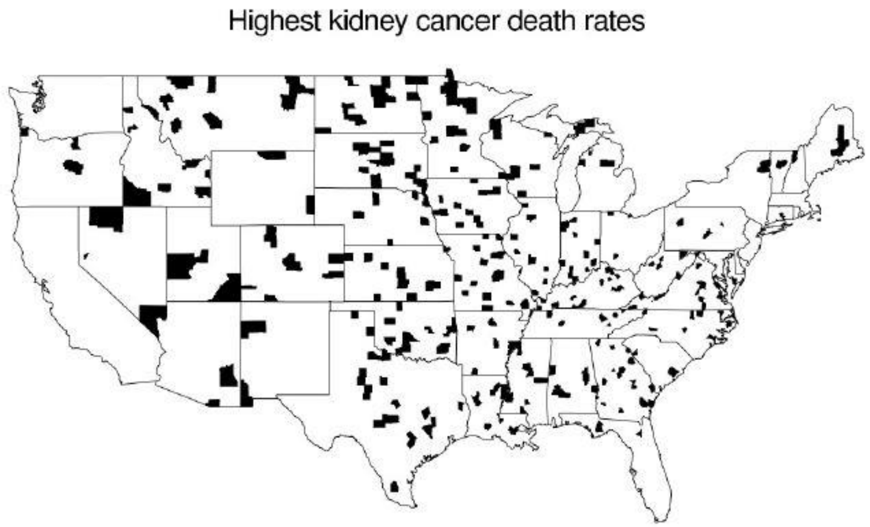
A sampling distribution of $y = (y_1, \ldots, y_n)$

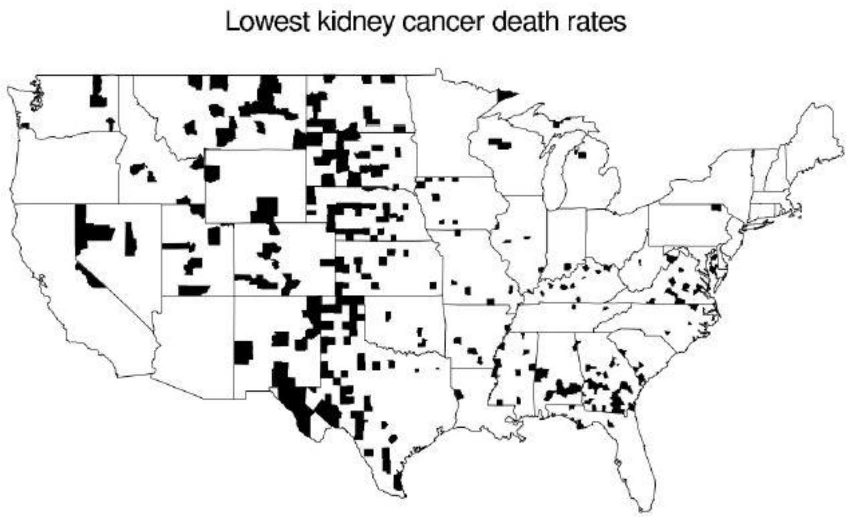
$$p(y|\theta) = \theta^n \exp(-n\overline{y}\theta), \quad \text{for } \overline{y} \geq 0$$

2.7 Example: informative prior distribution for cancer rates

Figure below shows the counties in US with the highest and lowest kidney caner death rates

→ Noticeably many cases are in the middle of the country





2.7 Example: informative prior distribution for cancer rates

- There might be some reason of this
- Perhaps sample size matters
- Example:

  Suppose a county A with population 1,000

  Since kidney cancer is a rare disease, A will have a high probability of 0 death case

  However, A still have a chance to have 1 case in 10 years, which will lead to put in the top 10% with ratio of 1 per 10,000 per year

2.7 Example: informative prior distribution for cancer rates

- Cancer death rates model

$$p(y_j|\theta_j) \sim \text{Poi}\left(10 n_j \theta_j\right)$$

- Notations

  y_j: # of kidney cancer deaths in county j

  n_j: population of the county

  θ_j: death rate per person per year

- For Bayesian inference,

  Need prior distribution for unknown rate θ_j

  Use Gamma distribution which is conjugate to the Poisson

  Consider an independent prior

  How about hyperparameters?

2.7 Example: informative prior distribution for cancer rates

- Constructing a prior distribution

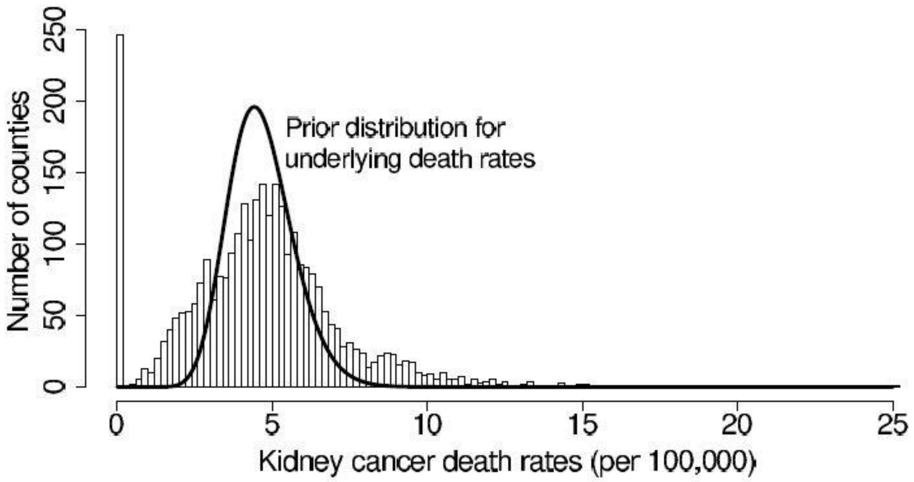
$$p(y_j) = \int p(y_j|\theta_j)P(\theta_j)\,\mathrm{d}\theta_j$$

Hence, $y_j \sim Negbin(\alpha, \frac{\beta}{10n_j})$ in this case

$$E(y_j) = 10n_j\frac{\alpha}{\beta} \quad \rightarrow \quad E(\frac{y_j}{10n_j}) = \frac{\alpha}{\beta}$$

$$\mathrm{var}(y_j) = 10n_j\frac{\alpha}{\beta} + (10n_j)^2\frac{\alpha}{\beta^2} \;\rightarrow\; \mathrm{var}\left(\frac{y_j}{10n_j}\right) = \frac{1}{10n_j}\frac{\alpha}{\beta} + \frac{\alpha}{\beta^2}$$

2.7 Example: informative prior distribution for cancer rates



- Posterior: Gamma distribution (conjugate to Poisson)

$$p(\theta_j|y_j) \sim Gamma(\alpha + y_j, \beta + n)$$

- Hyperparameter for $p(\theta_j|y_j)$

set $\alpha = 20,\ \beta = 430{,}000$

- Reasonable prior distribution for death rate in the U.S. during the period

2.7 Example: informative prior distribution for cancer rates

- Posterior distribution

$$p(\theta_j|y_j) \sim Gamma(20 + y_j, 430{,}000 + 10n_j)$$

$$\mathrm{E}(\theta_j|y_j) = \frac{20+y_j}{430{,}000+10n_j}$$

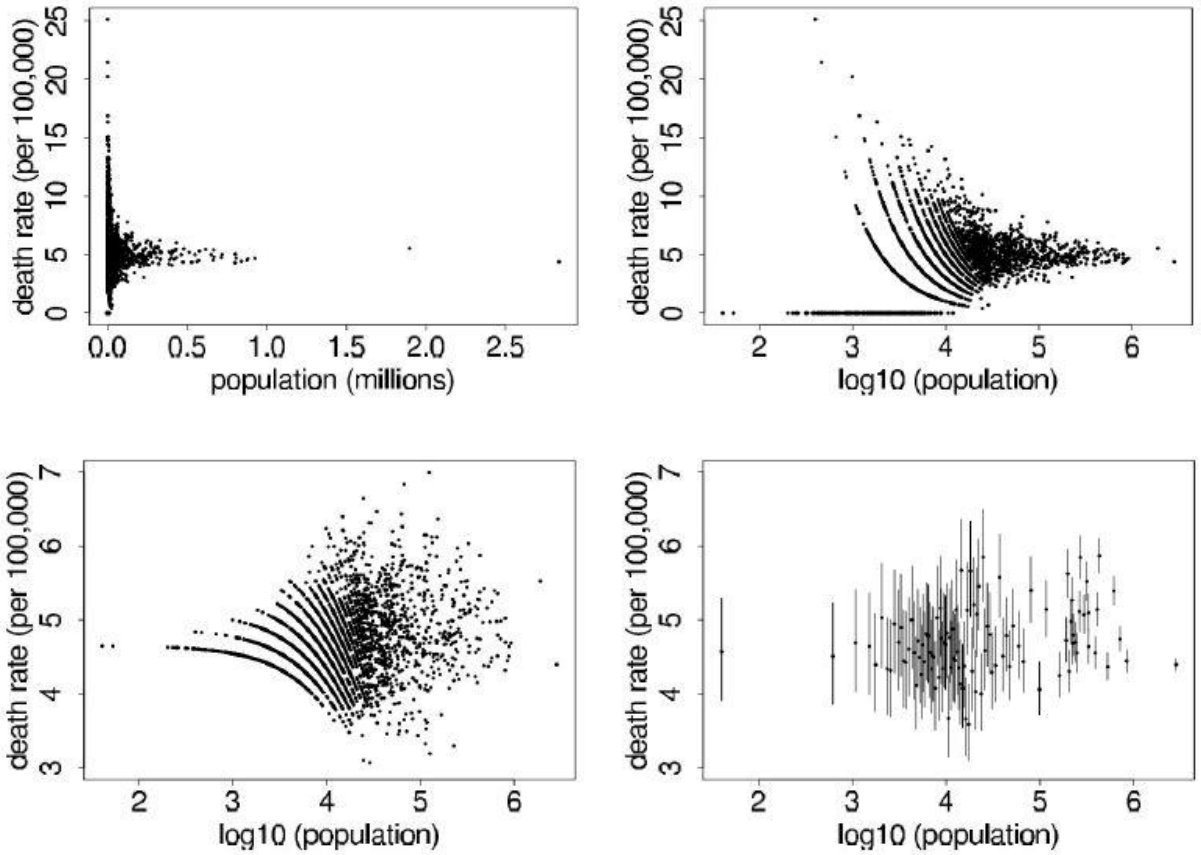
$$\mathrm{Var}(\theta_j|y_j) = \frac{20+y_j}{(430{,}000+10n_j)^2}$$

- For counties with small nj, the data are dominated by the prior.
- For counties with large nj, the data dominate the prior.

2.7 Example: informative prior distribution for cancer rates

- Comparing counties of different sizes

Bayes-estimated rates are much less variable



2.8 Noninformative prior distributions

- Desire for prior distributions to play a minimal role in the posterior distribution
- Noninformative prior shows vague information about the parameter
    - let the data speak for themselves
    - Diffuse or flat prior
    - Improper prior
    - Jeffrey's invariance principle
    - cf) weakly informative prior

2.8 Noninformative prior distributions

- Proper and improper prior distributions

Estimating mean θ of normal model with known variance σ^2

$$p(\theta|y) \sim N(\mu_0, \tau_0^2)$$

If $\tau_0^2 \to \infty$, the prior information(=$1/\tau_0^2$) vanishes and

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

If $P(\theta)$ is proportional to constant $\theta \in (-\infty, \infty)$, it is improper for this violates the assumption that probabilities sum to 1

$$\int P(\theta)\,\mathrm{d}\theta = \infty, and\ p(\theta|y) = N(\theta|\bar{y}, \sigma^2/n)$$

2.8 Noninformative prior distributions

- Improper prior can lead to proper posterior

$$\int P(\theta|y)\,\mathrm{d}\theta \; \alpha \int P(y|\theta)P(\theta)\,\mathrm{d}\theta < \infty$$

- Posterior distribution which is obtained from improper prior must be interpreted with great care!

2.8 Noninformative prior distributions

- Jeffrey's invariance principle

considering one-to-one transformations for the parameter

$$\phi = h(\theta)$$

By transformation of variables, $P(\theta)$ is equivalent to the following prior density on ϕ

$$P(\phi)p(\theta)\left|\frac{d\theta}{d\emptyset}\right| = P(\theta)|h'(\theta)|^{-1}$$

2.8 Noninformative prior distributions

- Jeffrey's invariance principle

This leads to defining the noninformative prior density as

$$P(\theta)\alpha[J(\theta)]^{1/2}$$

where $J(\theta)$ as the Fisher information for θ

$$J(\theta) = E\left(\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2\middle|\,\theta\right) = -E\left(\frac{\mathrm{d}^2\log_P(y|\theta)}{\mathrm{d}\theta^2}\middle|\,\theta\right)$$

2.8 Noninformative prior distributions

- Jeffrey's invariance principle

Jeffrey's prior is invariant to parametrization: For $\phi = h(\theta)$,

$$J(\phi) = -E\left(\frac{\mathrm{d}^2 \log P\,(y|\phi)}{\mathrm{d}\phi^2}\right)$$

$$= -E\left(\frac{\mathrm{d}^2 \log P\left(y|\theta = h^{-1}(\phi)\right)}{\mathrm{d}\theta^2}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|^2\right) = J(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|^2$$

Thus, $J(\phi)^{1/2} = J(\theta)^{1/2}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|$

2.8 Noninformative prior distributions

- Difficulties with noninformative prior distributions

  1. Searching for a prior distribution that is always vague seems misguided

  2. For many problems, there is no clear choice for a vague prior distribution, since a density that is flat or uniform in one parameterization will not be in another

  3. Further difficulties arise when averaging over a set of competing models that have improper prior distributions

2.9 Weakly informative prior distributions

- A prior distribution is proper but is set up so that the information it provides is intentionally weaker than actual prior knowledge is available
- In general any problem has some natural constraints that allow a weakly informative model
- Two principles for weakly informative priors

    Start with some version of noninformative prior and then add information

    Start with a informative prior and broaden it to account for uncertainty