Single Parameter Model

Anita Wang

2017 - 2018

Single parameter model

- Single parameter model is statistical models where only a single scalar parameter is to be estimated; that is, the estimand θ is **one-dimensional**

In this chapter:

- **Binomial**
- **Normal**
- **Poisson**
- **Exponential**

Binomial

- In the simple binomial model, the aim is to estimate an **unknown population proportion** from the results of a sequence of ‘Bernoulli trials’; that is, data $y_1, \ldots, y_n$.
- Because of the exchangeability, the data can be summarized by the total number of successes in the n trials, which we denote here by y.
- The binomial sampling distribution is

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

where on the left side we suppress the dependence on n because it is regarded as part of the experimental design that is considered fixed

Example

- We consider the estimation of the sex ratio within a population of human births. The currently accepted value of the proportion of female births in large European-race populations is 0.485.
- Let y be the number of girls in n recorded births. we are assuming that the n births are conditionally independent given θ, with the probability of a female birth equal to θ for all cases.
- For simplicity, we assume that the prior distribution for θ is uniform on the interval [0, 1].
- The posterior density,

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

Estimating the probability of a female birth

- Each of the four experiments has the same proportion of successes, but the sample sizes vary.

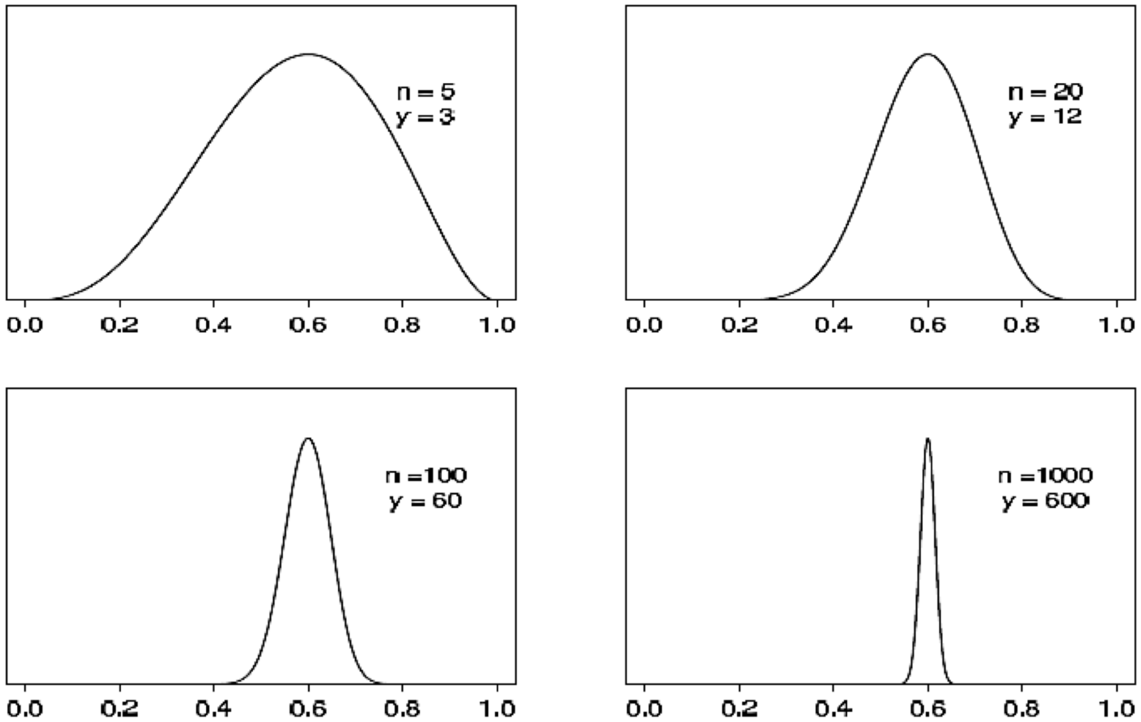


Figure 2.1 *Unnormalized posterior density for binomial parameter θ, based on uniform prior distribution and y successes out of n trials. Curves displayed for several values of n and y.*

Posterior distribution

- we can recognize the posterior distribution as the unnormalized form of the beta distribution

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

- The expectation of θ is

$$E(\theta) = (y + 1)/(n + 2)$$

- The probability that female birth proportion is larger than 0.5 is

$$\Pr(\theta > 0.5) = \int_{0.5}^{1} \text{Beta}(y + 1, n - y + 1)\, d\theta$$

$$= \int_{0.5}^{1} \frac{1}{B(y + 1, n - y + 1)} \theta^{y}(1 - \theta)^{n-y}\, d\theta$$

Prediction

- The prior predictive distribution is

$$p(y) = \int_0^1 \binom{n}{y} \theta^{y}(1-\theta)^{n-y}\, d\theta$$

$$= \int_0^1 \mathrm{Beta}(y+1, n-y+1)\frac{1}{n+1}\, d\theta = \frac{1}{n+1}, \text{ for } y = 0, \ldots, n.$$

Under the model, all possible values of y are equally likely.

- For posterior prediction from this model, we might be more interested in the outcome of one new trial, rather than another set of n new trials. Letting $\tilde{y}$ denote the result of a new trial, exchangeable with the first n,

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

Properties

- The prior mean of θ is the average of all possible posterior means over the distribution of possible data.

$$E(\theta) = E(E(\theta|y))$$

- The posterior variance is on average smaller than the prior variance, by an amount that depends on the variation in posterior means over the distribution of possible data

$$var(\theta) = E(var(\theta|y)) + var(E(\theta|y))$$

Properties

- In the binomial example with the uniform prior distribution, the prior mean is 1/2 , and the prior variance is 1/12 .
- The posterior mean, (y+1)/(n+2), is a **compromise** between the prior mean and the sample proportion, y/n,

- Conclusion:

The posterior distribution is centered at a point that represents a compromise between the prior information and the data, and the compromise is controlled to a greater extent by the data as the **sample size increases**.

Summarizing posterior inference

- For many practical purposes, various numerical summaries of the distribution are desirable.
- Commonly used summaries of location are the mean, median, and mode(s) of the distribution; variation is commonly summarized by the standard deviation, the interquartile range, and other quantiles.
- In the previous example,

$$E(\theta) = (y + 1)/(n + 2)$$

$$\text{Mode} = y/n$$

$$\sigma = \sqrt{\frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}}$$

Posterior quantiles and intervals

in the case of a $100(1-\alpha)\%$ interval, to the range of values above and below which lies exactly $100(\alpha/2)\%$ of the posterior probability.

$$\Pr(\theta < x) = \int_0^x \mathrm{Beta}(y+1, n-y+1)\, d\theta$$

$$= \int_0^x \frac{1}{B(y+1, n-y+1)} \theta^y (1-\theta)^{n-y}\, d\theta = \alpha/2$$

x can be obtained by R function or simulation

Informative prior distributions

- the prior distribution represents a population (or our knowledge) of possible parameter values, from which the θ of current interest has been drawn
- In the previous example, we use **uniform distribution** as a prior, which is **non-informative**, so that the prior predictive distribution for y (given n) is uniform on the discrete set $\{0, 1, \ldots, n\}$.
- A uniform specification is appropriate if nothing is known about θ.

Different prior densities

- We consider a parametric family of prior distributions that includes the uniform as a special case and construct a family of prior densities that lead to simple posterior densities.
- $\theta \sim \text{Beta}(\alpha, \beta)$:

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

- this prior density is equivalent to $\alpha - 1$ prior successes and $\beta - 1$ prior failures.
- The posterior density,

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

Conjugate prior

- The property that the posterior distribution follows the same parametric form as the prior distribution is called **conjugacy**; the beta prior distribution is a **conjugate family** for the binomial likelihood.
- If F is a class of sampling distributions $p(y|\theta)$, and P is a class of prior distributions for θ, then the class P is conjugate for F if

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

Posterior mean and variance

- the posterior mean of θ may be interpreted as the posterior probability of success for a future draw from the population:

$$E(\theta|y) = (\alpha + y)/(\alpha + \beta + n)$$

which always lies between the sample proportion, y/n, and the prior mean, $\alpha/(\alpha + \beta)$.

- The posterior variance is

$$\mathrm{var}(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E(\theta|y)[1 - E(\theta|y)]}{\alpha + \beta + n + 1}.$$

As y and $n - y$ become large with fixed α and β, $E(\theta|y) \approx y/n$ and $\mathrm{var}(\theta|y) \approx \frac{1}{n}\frac{y}{n}(1 - \frac{y}{n})$, which approaches zero at the rate $1/n$. In the limit, the parameters of the prior distribution have no influence on the posterior distribution.

Prediction

- The posterior prediction is

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

$$= \int \frac{(1-\theta)^{n-y+\beta-1}}{B(y+\alpha, n-y+\beta)}\theta^{y+\alpha-1}\binom{m}{\tilde{y}}\theta^{\tilde{y}}(1-\theta)^{\mathrm{m}-\tilde{y}}\,d\theta$$

$$= \binom{m}{\tilde{y}}\frac{B(y+\alpha+\tilde{y}, n-y+\beta+m-\tilde{y})}{B(y+\alpha, n-y+\beta)}$$

Normal mean with known variance: a single observation

- Consider a single scalar observation y from a normal distribution parameterized by a mean θ and variance σ^2, where for this initial development we assume that σ^2 is known.

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

The conjugate prior $\theta \sim N(\mu_0, \tau_0^2)$

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

hyperparameters μ_0 and τ_0^2.

Posterior distribution

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

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

where

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

- the posterior **precision** equals the prior precision plus the data precision.

Posterior distribution

- the posterior mean is expressed as a weighted average of the prior mean and the observed value, y, with weights proportional to the precisions.
- we can express μ_1 as the prior mean adjusted toward the observed y,

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

Or as the data 'shrunk' toward the prior mean,

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

- the posterior mean is a compromise between the prior mean

and the observed value.

Interpretation

- At the extremes, the posterior mean equals the prior mean or the observed data:

$$\mu_1 = \mu_0 \text{ if } y = \mu_0 \text{ or } \tau_0^2 = 0;$$
$$\mu_1 = y \text{ if } y = \mu_0 \text{ or } \sigma^2 = 0$$

If $\tau_0^2 = 0$, the prior distribution is infinitely more precise than the data, and so the posterior and prior distributions are identical and concentrated at the value μ_0.

If $\sigma^2 = 0$, the data are perfectly precise, and the posterior distribution is concentrated at the observed value y.

If $y = \mu_0$, the prior and data means coincide, and the posterior mean must also fall at this point.

Prediction

- Posterior predictive distribution:

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

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

The exponential of a quadratic function of $(\tilde{y}, \theta)$, hence $\tilde{y}$ and θ have a joint normal posterior distribution, and so the marginal posterior distribution of $\tilde{y}$ is normal.

Prediction

$$\tilde{y}|y \sim N(\mu_1, \sigma^2 + \tau_1^2)$$

Proof:

As $E(\tilde{y}|\theta) = \theta$ and $var(\tilde{y}|\theta) = \sigma^2$,

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

and

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

Thus, the posterior predictive distribution of $\tilde{y}$ has mean equal to the posterior mean of θ and two components of variance: the predictive variance σ^2 from the model and the variance τ_1^2 due to posterior uncertainty in θ.

Normal mean with known variance: more observations

- more realistic situation:

a sample of independent and identically distributed observations
y = $(y_1, \ldots, y_n)$ is available.

- Posterior density:

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta)p(y|\theta) \\
&= p(\theta)\prod_{i=1}^{n} p(y_i|\theta) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}(\theta-\mu_0)^2+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\right).
\end{aligned}
$$

Posterior distribution

The posterior distribution is also a normal distribution:

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

where

$$\mu_n = \frac{\frac{\mu_0}{\tau_0^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}, \qquad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

Incidentally, the same result is obtained by adding information for the data $y_1, \ldots, y_n$ one point at a time, using the posterior distribution at each step as the prior distribution for the next

Property

- the prior precision, $1/\tau_0^2$, and the data precision, n/σ^2, play equivalent roles, so if n is large, the posterior distribution is largely determined by σ^2 and the sample value $\bar{y}$.

- For example, if $\tau_0^2 = \sigma^2$, then the prior distribution has the same weight as one extra observation with the value μ_0.

- More specifically, as $\tau_0 \to \infty$ with n fixed, or as $n \to \infty$ with τ_0^2 fixed, we have:

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

Normal distribution with known mean but unknown variance

- For $p(y|\theta, \sigma^2) = N(y|\theta, \sigma^2)$, with θ known and σ^2 unknown, the likelihood for a vector y of n i.i.d observations is

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

$$= (\sigma^2)^{-n/2} \exp(-\frac{n}{2\sigma^2}v)$$

The sufficient statistics is

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

Prior

- The conjugate prior density is a scaled inverse-χ^2 distribution with scale σ_0^2 and degrees of freedom ν_0.

$$p(\sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}+1} \exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)$$

- Posterior density

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

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

Inverse Chi-square distribution

- Chi-square Distribution ($\chi_\nu^2 = Gamma(\frac{\nu}{2}, \frac{1}{2})$)

$$p(y|\nu) = \frac{2^{\nu/2}}{\Gamma(\nu/2)} y^{\frac{\nu}{2}-1} e^{-\frac{y}{2}}$$

$$E(y) = \nu \qquad \mathrm{Var}(y) = 2\nu$$

- Inverse Chi-square ($\mathrm{Inv} - \chi_\nu^2$)

$$y \sim \mathrm{Inv} - \chi_\nu^2 \text{ if } \frac{1}{y} \sim \chi_\nu^2$$

Note that $\mathrm{Inv} - \chi_\nu^2 = \mathrm{Inv} - \mathrm{gamma}(\frac{\nu}{2}, \frac{1}{2})$

$$p(y|\nu) = \frac{2^{-\nu/2}}{\Gamma(\nu/2)} y^{-(\frac{\nu}{2}+1)} e^{-\frac{1}{2y}}$$

$$E(y) = \frac{1}{\nu - 2} \qquad \mathrm{Var}(y) = \frac{2}{(\nu-2)^2(\nu-4)}$$

Scaled Inverse Chi-square

- Scaled Inverse Chi-square ($\mathrm{Inv}-\chi^2(\nu, s^2)$)

$y \sim \mathrm{Inv}-\chi^2(\nu, s^2)$ if $\frac{\nu s^2}{y} \sim \chi^2_\nu$

- Note that $\mathrm{Inv}-\chi^2(\nu, s^2) = \mathrm{Inv}-\mathrm{gamma}(\frac{\nu}{2}, \frac{\nu}{2} s^2)$

$$p(y|\nu) = \frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}}{\Gamma(\nu/2)} s^\nu y^{-(\frac{\nu}{2}+1)} e^{-\frac{\nu s^2}{2y}}$$

CDF: $P_{I\chi^2}(y, \nu, s^2) = 1 - P_{\chi^2}(\frac{\nu s^2}{y}, \nu)$

Quantile Function: $P^{-1}_{I\chi^2}(p, \nu) = \frac{\nu s^2}{P^{-1}_{\chi^2}(1-p, \nu)}$

Scaled Inverse Chi-square

$$E(y) = \frac{\nu}{\nu - 2} s^2$$

$$\mathrm{Var(y)} = \frac{2\nu^2}{(\nu - 2)^2(\nu - 4)} s^4$$

$$\mathrm{Mode}(y) = \frac{\nu}{\nu + 2} s^2$$

Exercise

- The posterior density of σ^2 is

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

Calculate the posterior expectation and variance

Poisson distribution

- Observations: $y = (y_1, y_2, \ldots, y_n)$
- Likelihood:

$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \propto \theta^{n\bar{y}} e^{-n\theta}$$

- Prior density: Gamma(α, β)

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

- Posterior density:

$$p(\theta|y) \propto e^{-(n+\beta)\theta} \theta^{n\bar{y}+\alpha-1}$$

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

Pediction

- Posterior density:

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

- the Poisson model for a single observation, y, has prior predictive distribution

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

known as the negative binomial density

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

The negative binomial is a robust alternative to the Poisson distribution

Poisson model parameterized in terms of rate and exposure

- In many applications, it is convenient to extend the Poisson model for data points $y_1, \ldots, y_n$ to the form

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

  - the values x_i are known positive values of an explanatory variable x (called the **exposure** of the ith unit)
  - θ is the unknown parameter of interest (called **rate**)
- Likelihood

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

- Posterior density

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

Example

- Suppose that causes of death are reviewed in detail for a city in the United States for a single year. It is found that 3 persons, out of a population of 200,000, died of asthma, giving a crude estimated asthma mortality rate in the city of 1.5 cases per 100,000 persons per year.
- under the Poisson model, the sampling distribution of y may be expressed as Poisson(2.0θ), where θ represents the true underlying long-term asthma mortality rate in our city (measured in cases per 100,000 persons per year).
- $y = 3$ is a single observation with exposure $x = 2.0$ and unknown rate θ.

Prior distribution

Reviews of asthma mortality rates around the world suggest that mortality rates above 1.5 per 100,000 people are rare in Western countries, with typical asthma mortality rates around **0.6** per 100,000.

Trial-and-error exploration reveals that a Gamma(3.0, 5.0) density provides a plausible prior density with the mean of this prior distribution is **0.6** (with a mode of 0.4), and 97.5% of the mass of the density lies below 1.44 .

Posterior distribution

- The posterior distribution is Gamma($\alpha + y$, $\beta + x$), which is **Gamma(6.0,7.0),** has mean 0.86—substantial shrinkage has occurred toward the prior distribution. The death rate is more than 1.0 per 100,000 per year is 0.30.
- To consider the effect of additional data, suppose that ten years of data are obtained for the city in our example, we find $y = 30$ deaths over 10 years. The posterior distribution of θ is then **Gamma(33.0, 25.0)**. After ten years of data, the posterior mean of θ is 1.32, and the posterior probability that θ exceeds 1.0 is 0.93.

Posterior density plot

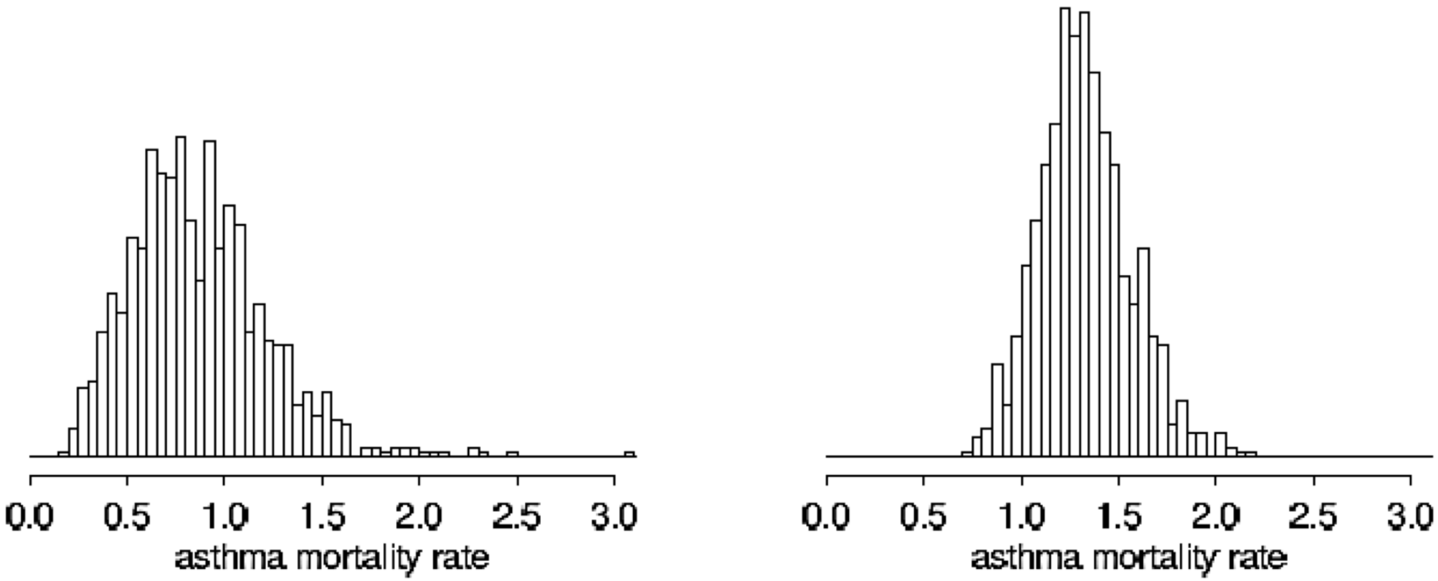


Figure 2.5 *Posterior density for θ, the asthma mortality rate in cases per 100,000 persons per year, with a Gamma$(3.0, 5.0)$ prior distribution: (a) given $y = 3$ deaths out of 200,000 persons; (b) given $y = 30$ deaths in 10 years for a constant population of 200,000. The histograms appear jagged because they are constructed from only 1000 random draws from the posterior distribution in each case.*

Exponential Distribution

- Observations: $y = (y_1, y_2, \dots, y_n)$
- Likelihood:

$$p(y|\theta) = \theta^n \exp(-n\bar{y}\theta)$$

- Prior density: Gamma(α, β)

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

- Posterior density:

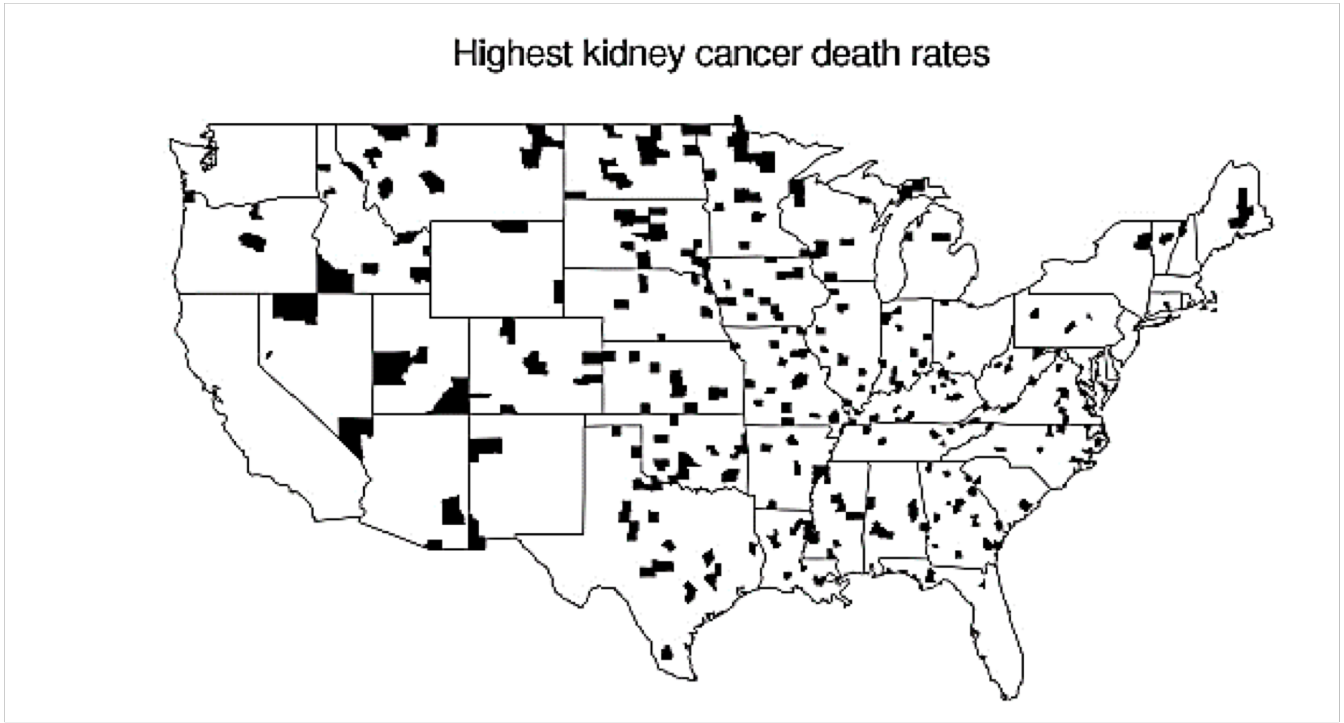
$$p(\theta|y) \propto \theta^{\alpha+n-1} \exp(-(n\bar{y} + \beta)\theta)$$

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

The sampling distribution when viewed as the likelihood of θ, for fixed y, is proportional to a Gamma(n+1, ny) density. Thus the Gamma(α, β) prior distribution for θ can be viewed as α−1 exponential observations with total waiting time β

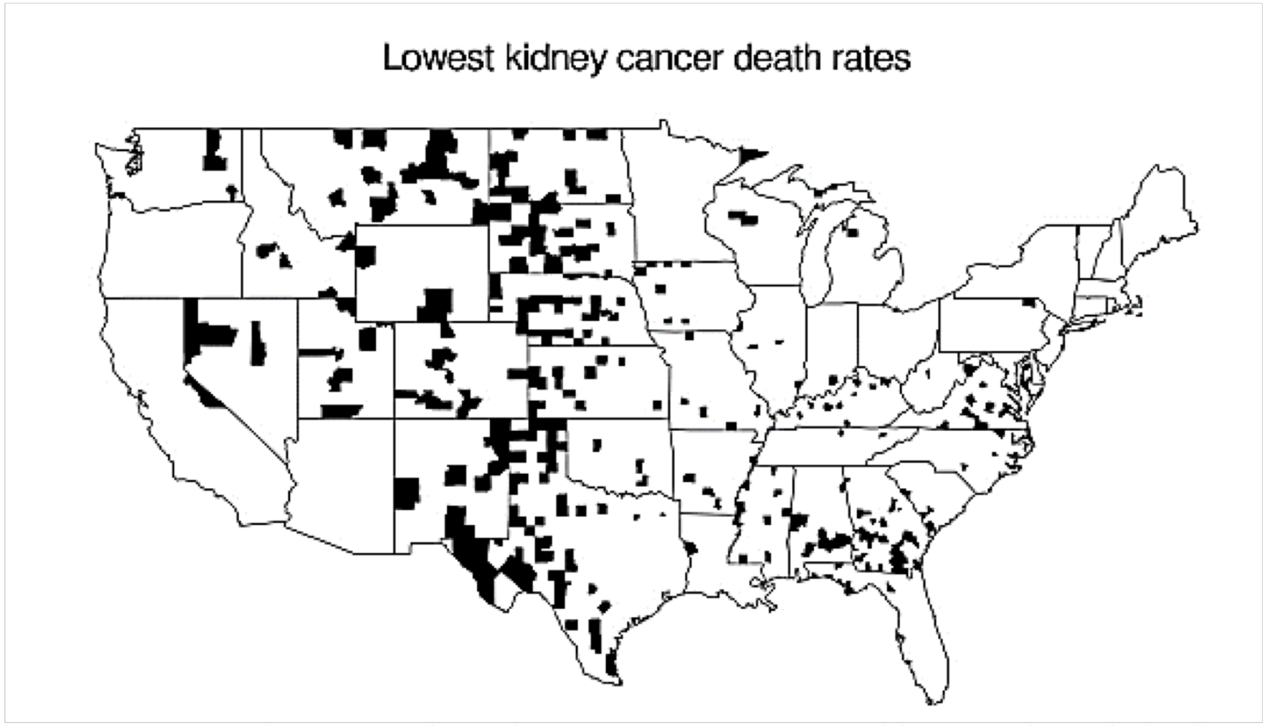
Example

- Figure 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.



A puzzling pattern in a map

- 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 . . . **Not the reason**



- why these areas have the lowest, as well as the highest, rates

A puzzling pattern in a map

- 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

- There is a chance 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.
- The Great Plains has many low-population counties, and so it is overrepresented in both maps.

Prior density

- Suppose the sampling data

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

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.

- The above figures are plotting $\frac{y_j}{10 n_j}$

- Prior distribution

$$\theta_j \sim \text{Gamma}(20\ , 430{,}000),$$

with mean $\alpha/\beta = 4.65 \times 10^{-5}$ and standard deviation $\sqrt{\alpha}/\beta = 1.04 \times 10^{-5}$.

Inference for a small county

- The posterior distribution of θ_j is then,

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

- The mean and variance

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

- Inference of a small county with $n_j = 1000$
  - 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 10^{-4} per person-year but the posterior mean is only $21/440{,}000 = 4.77 \times 10^{-5}$.

Inference for a large county

- Now consider a large county with n_j = 1 million.
- Likelihood: Poisson(10^7 θj)
- Prior distribution: Gamma(20, 430,000)
- Posterior distribution: Gamma(20 + y_j , 430,000 + 10^7)
  - If y_j = 393, then the raw death rate is 3.93×10^{-5} and the posterior mean of θ_j is $\frac{20+393}{10^7+430{,}000} = 3.96 \times 10^{-5}$,
  - 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.

Example continued

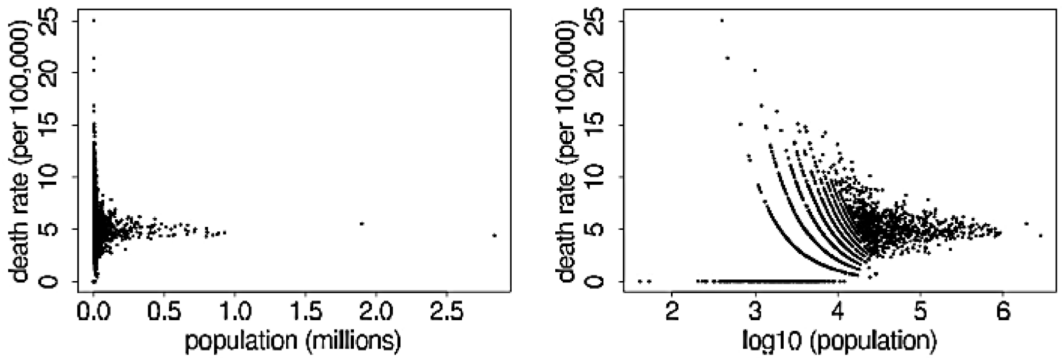


Figure 2.8 *(a) Kidney cancer death rates $y_j/(10n_j)$ vs. population size n_j. (b) Replotted on the scale of* $\log_{10}$ *population to see the data more clearly. The patterns come from the discreteness of the data* $(n_j = 0, 1, 2, \ldots)$.

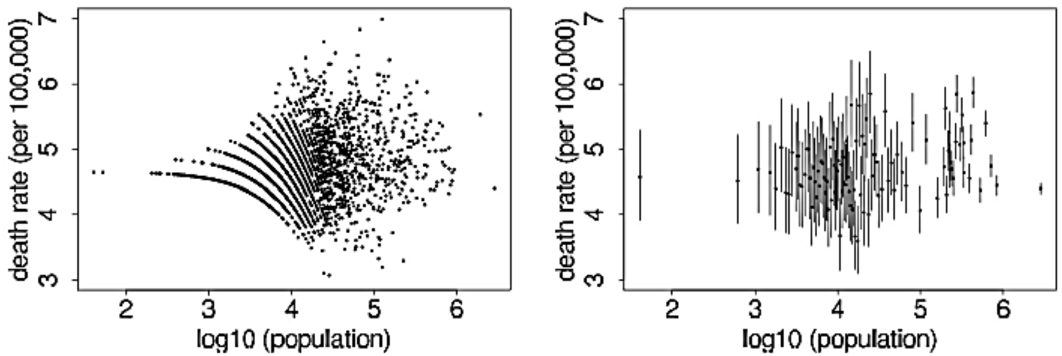


Figure 2.9 *(a) Bayes-estimated posterior mean kidney cancer death rates,* $E(\theta_j|y_j) = \frac{20+y_j}{430{,}000+10n_j}$ *vs. logarithm of population size n_j, the 3071 counties in the U.S. (b) Posterior medians and 50% intervals for θ_j for a sample of 100 counties j. The scales on the y-axes differ from the plots in Figure 2.8b.*

Thanks