Bayesian Statistics

Anita Wang

2017 - 2018

Bayes' Rule

The **joint** probability mass or density function can be written as a product of the **prior distribution** $p(\theta)$ and **the sampling distribution** $p(y|\theta)$

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Bayes' Rule:

The **posterior** density

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

where $p(y) = \sum_{\theta} p(\theta)p(y|\theta)$ over all possible value of θ (or $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous θ).

Bayes' Rule

An equivalent form omits the factor p(y), which does not depend on θ and, with fixed y, can thus be considered a constant, yielding the **unnormalized posterior density**,

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

The second term in this expression, p(y|θ), is taken here as a function of θ, not of y.

Prediction

- **Prior predictive distribution** (also called marginal distribution of y)

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

prior because it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable.

Prediction

- Posterior predictive distribution

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

Once the data y have been observed, the unknown observable $\tilde{y}$ can be predicted. For example, $y = (y_1, y_2, \ldots, y_n)$ may be the vector of recorded weights of an object weighed n times on a scale, $\theta = (\mu, \sigma^2)$ is the prior, and $\tilde{y}$ may be the yet to be recorded weight of the object in a planned new weighing.

Likelihood

Using Bayes' rule with a chosen probability model means that the data y affect the posterior inference only through $p(y|\theta)$. $p(y|\theta)$ is regarded as a function of θ for fixed y is **likelihood function**.

Likelihood principle

for a given sample of data, any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ.

Likelihood and odds ratios

- The ratio of the posterior density $p(\theta|y)$ evaluated at the points θ_1 and θ_2 under a given model is called the posterior odds for θ_1 compared to θ_2.

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)}$$

The posterior odds are equal to the prior odds multiplied by the likelihood ratio $p(y|\theta_1)/p(y|\theta_2)$.

Example

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her two X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes, and this is rare, since the frequency of occurrence of the gene is low in human populations.

Prior distribution

- Consider a woman who has an affected brother, which implies that her mother must be a carrier of the hemophilia gene with one 'good' and one 'bad' hemophilia gene. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene.
- Set: to be a carrier of the gene ($\theta = 1$) or not ($\theta = 0$)

- The prior distribution for unknown θ is

$$\Pr(\theta = 1) = \Pr(\theta = 0) = \frac{1}{2}$$

Likelihood

- The data used to update the prior information consist of the affection status of the woman's sons. Suppose she has two sons, neither of whom is affected. The outcomes of the two sons are exchangeable independent
- Let $y_i = 1$ or 0 denote an affected or unaffected son.

- The likelihood function is:

$$\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 1) = (0.5)(0.5) = 0.25$$

$$\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 0) = (1)(1) = 1$$

Posterior distribution

- In particular, interest is likely to focus on the posterior probability that the woman is a carrier.

$$\begin{aligned}\Pr(\theta = 1|y) &= \frac{p(y|\theta = 1)\Pr(\theta = 1)}{p(y|\theta = 1)\Pr(\theta = 1) + p(y|\theta = 0)\Pr(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.20.\end{aligned}$$

- Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes' rule provides a formal mechanism for determining the extent of the correction.

Prior and Posterior Odds

- The prior odds of the woman being a carrier

$$\frac{p(\theta = 1)}{p(\theta = 0)} = 0.5/0.5 = 1$$

- The likelihood ratio is

$$\frac{\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 1)}{\Pr(y_1 = 0,\ y_2 = 0 \mid \theta = 0)} = 0.25/1 = 0.25$$

- The posterior odds is $1(0.25) = 0.25$
- Conversely, the posterior probability

$$p(\theta = 1|y) = 0.25/(1+0.25) = 0.2$$

Adding more data

- A key aspect of Bayesian analysis is the ease with which sequential analyses can be performed.
- For example, suppose that the woman has a third son, who is also unaffected. The entire calculation does not need to be redone; rather we use the previous posterior distribution as the new prior distribution

$$\Pr(\theta = 1 | y_1, y_2, y_3) = \frac{(0.5)(0.2)}{(0.5)(0.2) + (1)(0.8)} = 0.111$$

Example

- Spelling correction is an important technique. Suppose someone types ‘radom.’ How should that be read? It could be a misspelling or mistyping of ‘random’ or ‘radon’ or some other alternative, or it could be the intentional typing of ‘radom’ (as in its first use in this paragraph).

$$\Pr(\theta|y = \text{‘radom’}) \propto p(\theta)\Pr(y = \text{‘radom’}|\theta)$$

$$p(\text{random}|\ \text{‘radom’}) = \frac{p(\theta_1)p(\text{‘radom’}|\theta_1)}{\sum_{j=1}^{3} p(\theta_j)p(\text{‘radom’}|\theta_j)}$$

$$\theta_1 = \text{random},\ \theta_2 = \text{radon},\ \theta_3 = \text{radom}$$

Prior distribution

- The prior probabilities $p(\theta_j)$ can most simply come from frequencies of these words in some large database

θ	$p(\theta)$
random	7.60×10^{-5}
radon	6.05×10^{-6}
radom	3.12×10^{-7}

- For the documents that we encounter, the relative probability of 'radom' seems much too high. We looked up the word in Wikipedia and found that it is a medium-sized city, home to 'the largest and best-attended air show in Poland . . .'
- We may have prior information or beliefs that have not yet been included in the model.

Likelihood

- Here are some conditional probabilities from Google's model of spelling and typing errors:

θ	$p(\text{'radom'}\|\theta)$
random	0.00193
radon	0.000143
radom	0.975

- 97% chance that this particular five-letter word will be typed correctly, a 0.2% chance of obtaining this character string by mistakenly dropping a letter from 'random,' and a much lower chance of obtaining it by mistyping the final letter of 'radon.'

Posterior distribution

- We multiply the prior probability and the likelihood to get joint probabilities and then renormalize to get posterior probabilities:

θ	$p(\theta)p(\text{`radom'}\|\theta)$	$p(\theta\|\text{`radom'})$
random	1.47×10^{-7}	0.325
radon	8.65×10^{-10}	0.002
radom	3.04×10^{-7}	0.673

- the typed word ‘radom’ is about twice as likely to be correct as to be a typographical error for ‘random,’ and it is very unlikely to be a mistaken instance of ‘radon.’

Decision making, model checking and improvement

- The first approach is to accept that the word was typed correctly
- The second option would be to question this probability by saying, for example, that ‘radom’ looks like a typo and that the estimated probability of it being correct seems much too high.
- Questioning the prior: The prior probabilities, on the other hand, are highly context dependent. The word ‘random’ is of course highly frequent in our own writing on statistics, ‘radon’ occurs occasionally, while ‘radom’ was entirely new to us.
- Label x as the contextual information

$$p(\theta|x, y) \propto p(\theta|x)p(y|\theta, x).$$

Thanks

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.

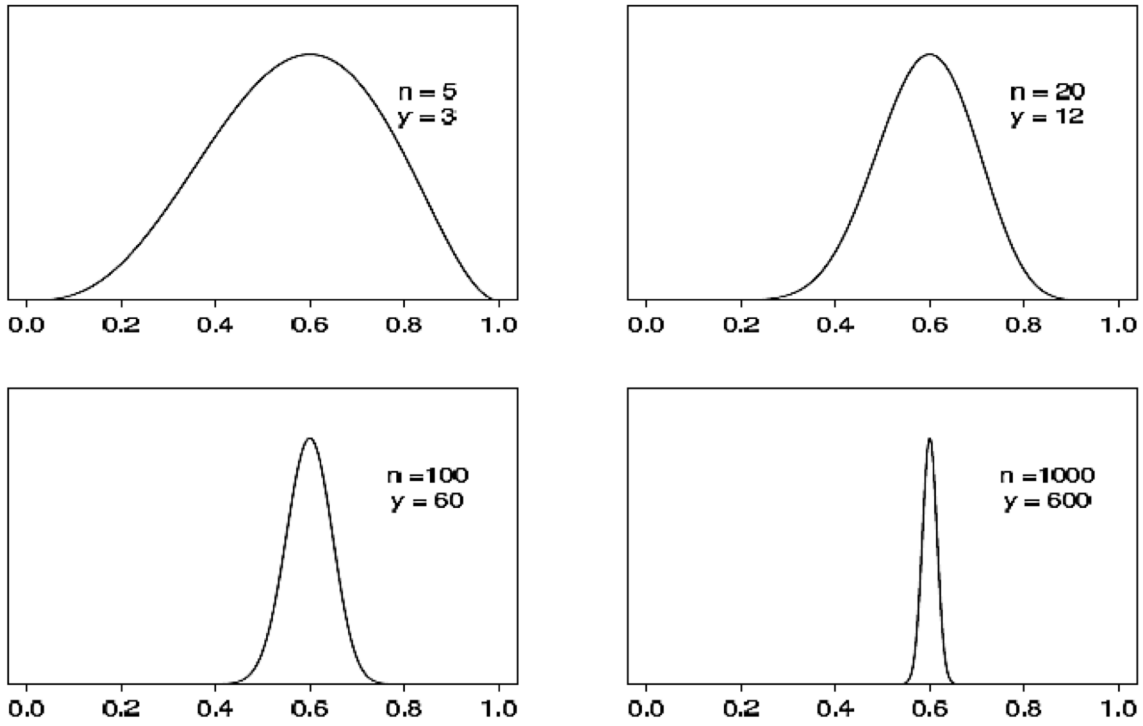


Figure 2.1 *Unnormalized posterior density for binomial parameter θ, based on uniform prior distribution and y successes out of n trials. Curves displayed for several values of n and y.*

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

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

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

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.

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

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

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

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

As y and $n - y$ become large with fixed α and β, $E(\theta|y) \approx y/n$ and $\text{var}(\theta|y) \approx \frac{1}{n}\frac{y}{n}(1 - \frac{y}{n})$, 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

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

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^2distribution 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 ($\chi_v^2 = Gamma(\frac{v}{2}, \frac{1}{2})$)

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

$$E(y) = v \qquad \text{Var}(y) = 2v$$

- Inverse Chi-square ($\text{Inv} - \chi_v^2$)

$$y \sim \text{Inv} - \chi_v^2 \text{ if } \frac{1}{y} \sim \chi_v^2$$

Note that $\text{Inv} - \chi_v^2 = \text{Inv} - \text{gamma}(\frac{v}{2}, \frac{1}{2})$

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

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

Scaled Inverse Chi-square

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

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

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

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

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

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

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

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

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

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

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

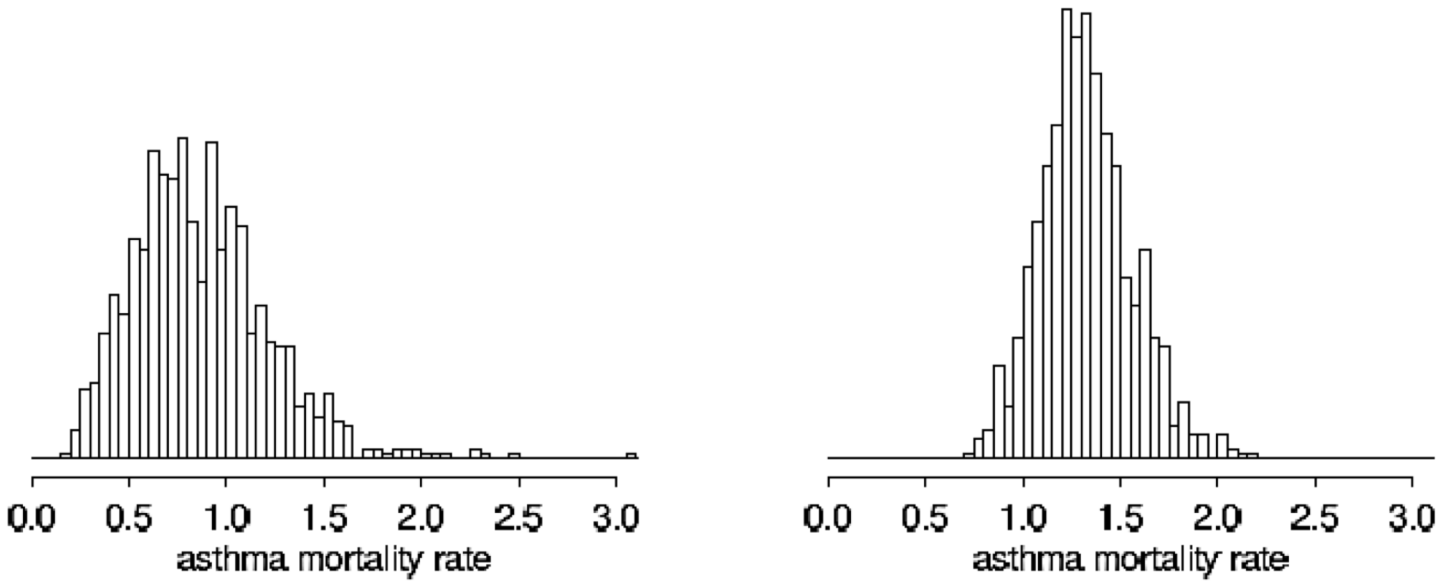


Figure 2.5 *Posterior density for θ, the asthma mortality rate in cases per* 100,000 *persons per year, with a Gamma*$(3.0, 5.0)$ *prior distribution: (a) given* $y = 3$ *deaths out of* 200,000 *persons; (b) given* $y = 30$ *deaths in* 10 *years for a constant population of* 200,000. *The histograms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

Exponential Distribution

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

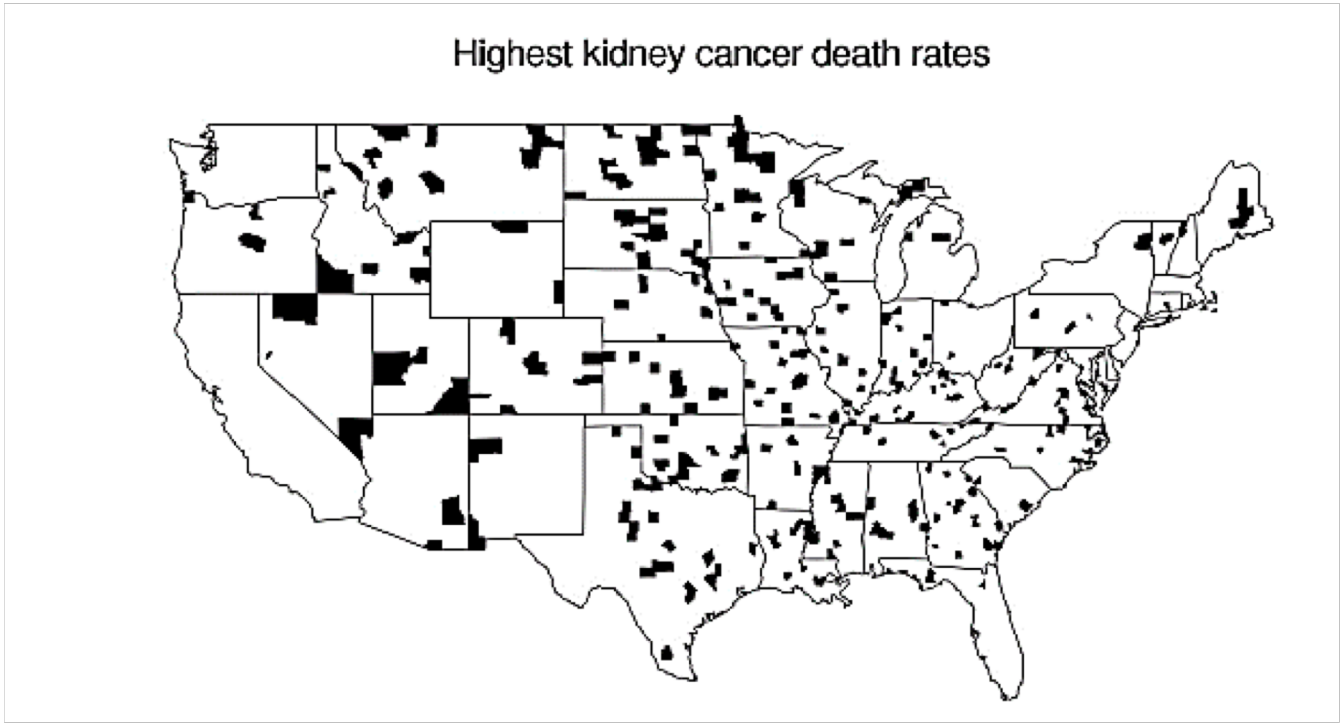
$$p(\theta|y) \propto \theta^{\alpha+n-1} \exp(-(n\bar{y} + \beta)\theta)$$

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

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

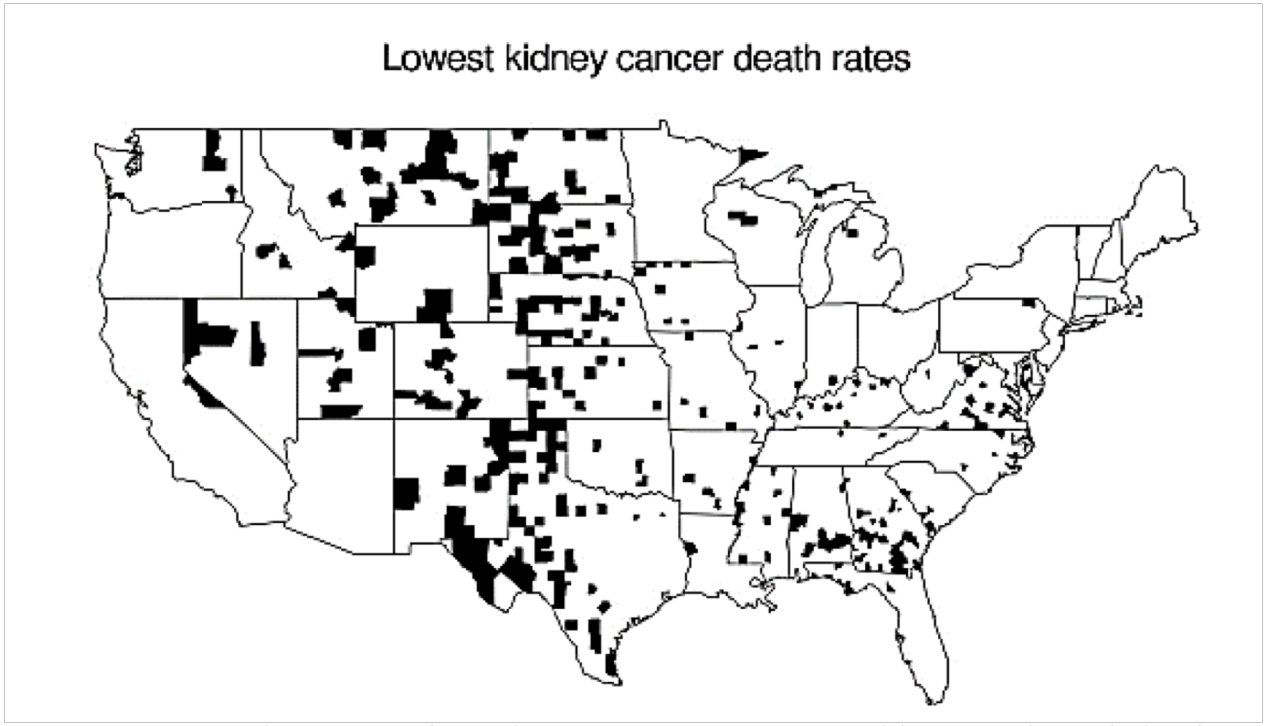
Example

- 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

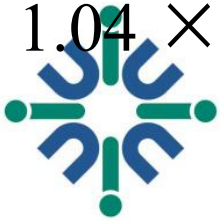
$$y_j \sim \text{Poisson}(10 n_j \theta_j),$$

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

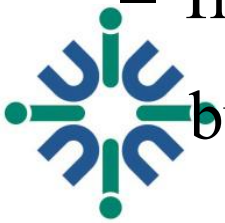
- The posterior distribution of θ_j is then,

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

- The mean and variance

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

- 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

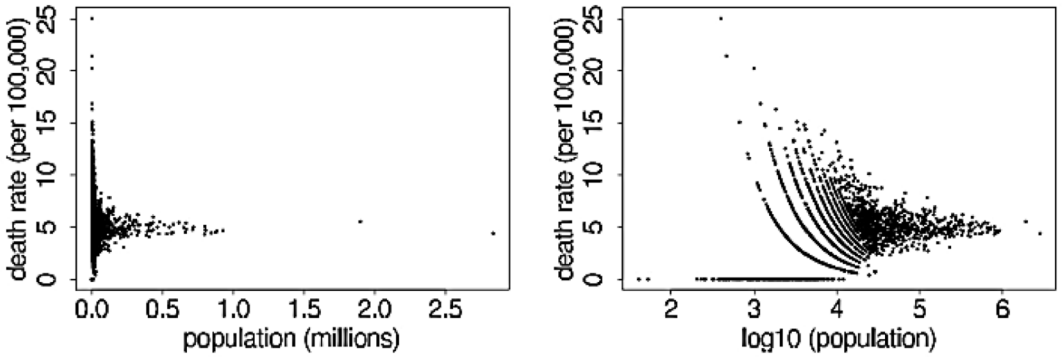


Figure 2.8 *(a) Kidney cancer death rates $y_j/(10n_j)$ vs. population size n_j. (b) Replotted on the scale of* $\log_{10}$ *population to see the data more clearly. The patterns come from the discreteness of the data* ($n_j = 0, 1, 2, \ldots$).

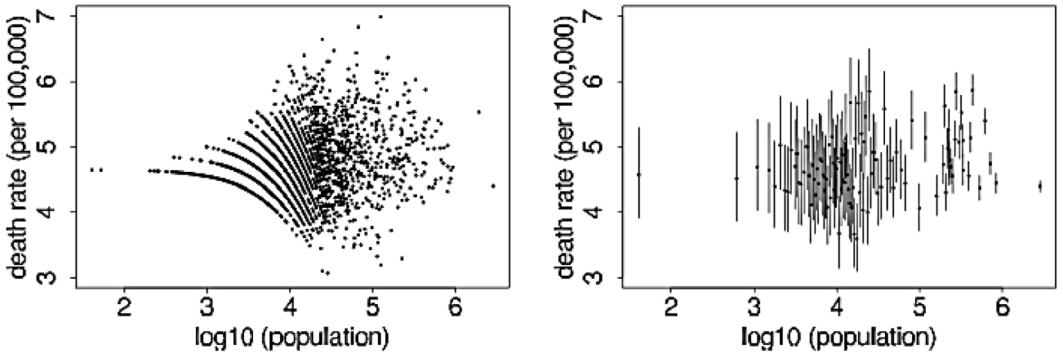


Figure 2.9 *(a) Bayes-estimated posterior mean kidney cancer death rates,* $E(\theta_j|y_j) = \frac{20+y_j}{430\,000+10n_j}$ *vs. logarithm of population size n_j, the 3071 counties in the U.S. (b) Posterior medians and 50% intervals for θ_j for a sample of 100 counties j. The scales on the y-axes differ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

Thanks

Prior

Anita Wang

2017 - 2018

Binomial example with different prior distributions

- In the binomial example, we have considered the uniform prior distribution for θ that the prior predictive distribution for y (given n) is uniform on the discrete set $\{0, 1, \ldots, n\}$.
- We can also parameterize a prior density as $\theta \sim \text{Beta}(\alpha, \beta)$:

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

As the likelihood is

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

Comparing $p(\theta)$ and $p(y|\theta)$ suggests that this prior density is equivalent to $\alpha - 1$ prior successes and $\beta - 1$ prior failures.

conjugacy

- The parameters of the prior distribution are often referred to as **hyperparameters,** which means we can specify a particular prior distribution by fixing two features of the distribution, for example its mean and variance.
- The posterior density for θ is

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

- 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.

Conjugate prior

- The conjugate family is mathematically convenient in that the posterior distribution follows a known parametric form.
- If information is available that contradicts the conjugate parametric family it may be necessary to use a more realistic prior distribution.
- Definition: 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

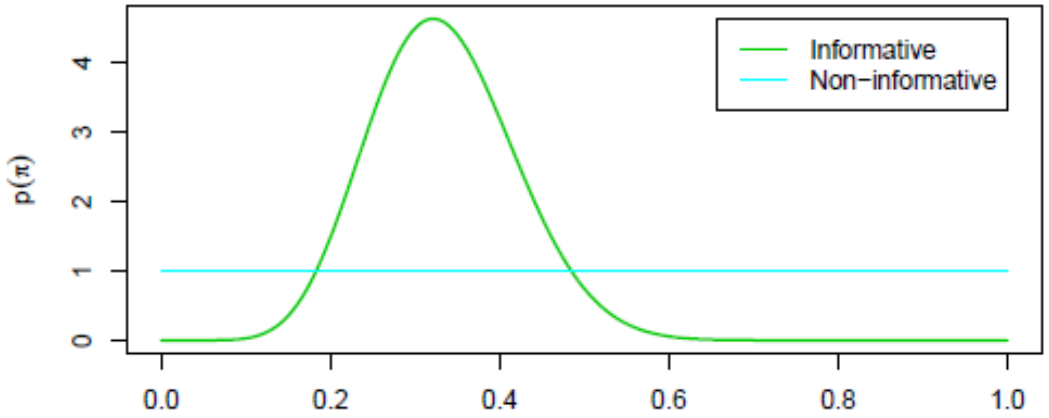
$$p(\theta|y) \in P \text{ for all } p(\cdot|\theta) \in F \text{ and } p(\cdot) \in P.$$

- Conjugate prior distributions have the practical advantage, in addition to computational convenience, of being interpretable as **additional data**.

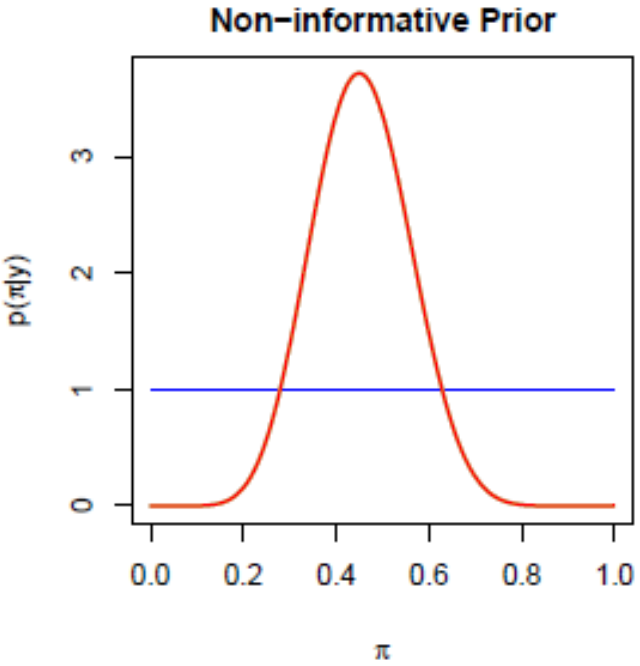
Noninformative prior distributions

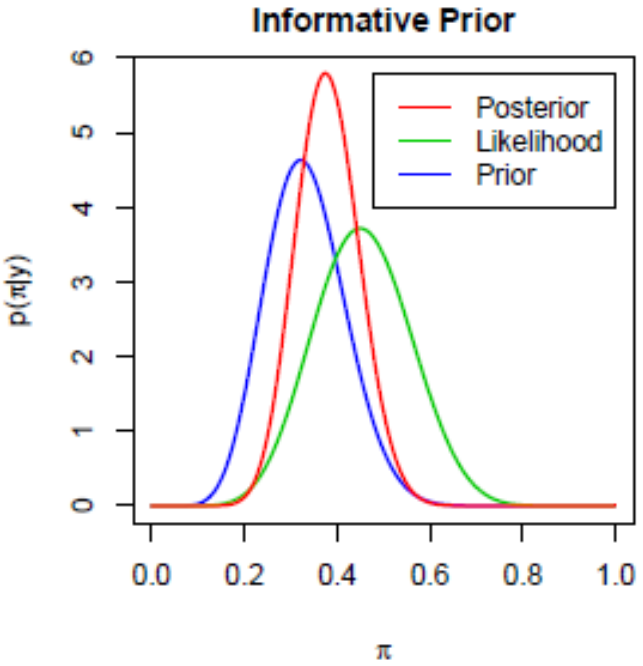
- When prior distributions have no population basis, they can be difficult to construct
- There has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution
- Such distributions are sometimes called ‘reference prior distributions,’ and the 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.

Informative vs Non-informative



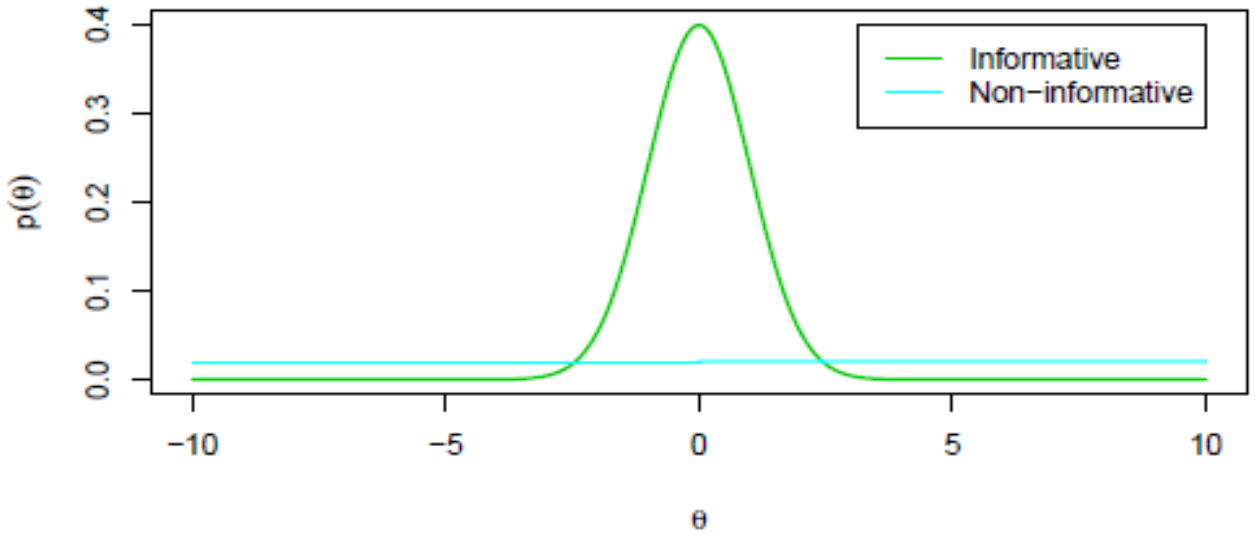
For this example, with the non-informative prior, Posterior=Likelihood





Improper prior distributions

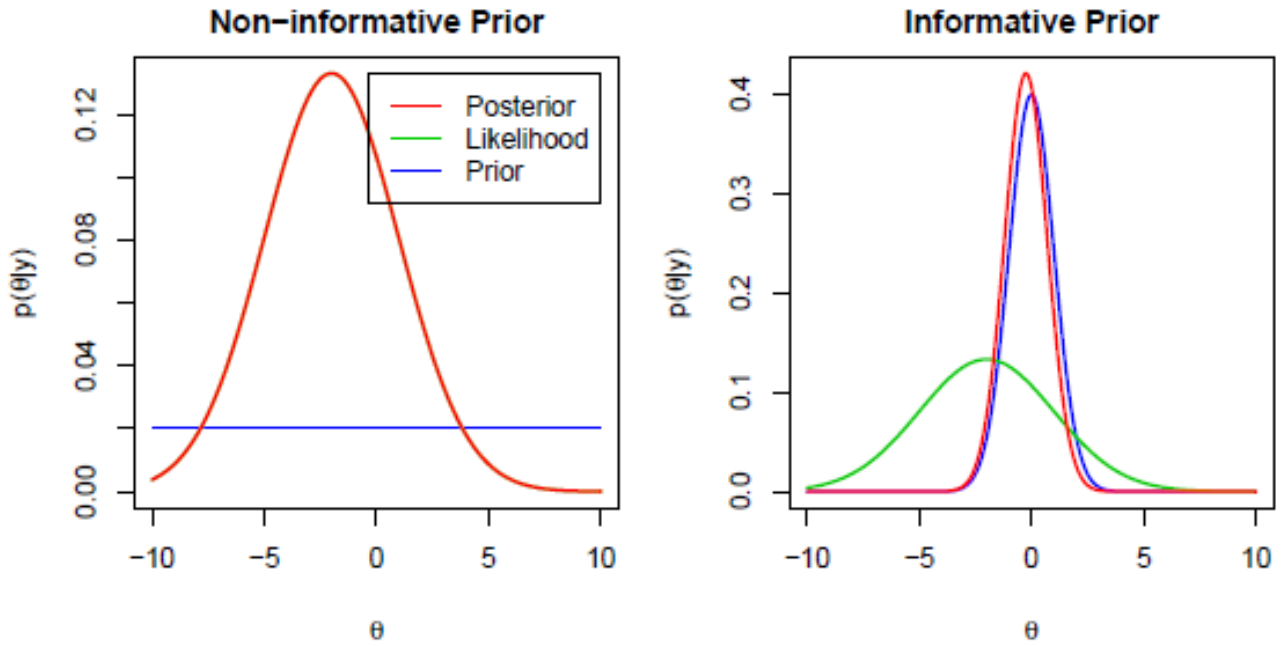
- However for a parameter that occurs on an infinite interval (e.g. a normal mean μ), using a uniform prior on μ is problematic.
- For the normal mean example, lets use the conjugate prior $N(\mu_0, \tau_0^2)$, but with a very big variance τ_0^2



Non-informative Prior

- The posterior mean and precision are

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



Non-informative Prior

- So if we let $\tau_0^2 \to \infty$, then

$$\mu_n \to \bar{y} \text{ and } \frac{1}{\tau_n^2} \to \frac{n}{\sigma^2}$$

- This equivalent to the posterior being proportional to the likelihood, which is what we get if $p(\theta) \propto 1$(e.g. uniform).

- This does not describe a valid probability density as

$$\int_{-\infty}^{\infty} d\theta = \infty$$

Proper

- A prior is called proper if it is a valid probability distribution

$$p(\theta) \geq 0, \forall \theta \in \Theta \text{ and } \int p(\theta)\, d\theta = 1$$

(If p(θ) integrates to any positive finite value, it is called an **unnormalized density** and can be renormalized—multiplied by a constant—to integrate to 1.

If a prior is proper, so must the posterior.

Improper

- A prior is called improper if

$$p(\theta) \geq 0, \forall \theta \in \Theta \text{ and } \int p(\theta)\, d\theta = \infty$$

- The prior distribution is improper in this example, but the posterior distribution is proper i.e., $\int p(\theta|y)d\theta$ is finite for all y, given at least one data point.
- Posterior distributions obtained from improper prior distributions must be interpreted with great care—one must always check that the posterior distribution has a finite integral and a sensible form.

Another example of noninformative prior

- The normal model with known mean but unknown variance

- The prior degrees of freedom, $\nu_0 \rightarrow 0$

$$p(\sigma^2|y) \approx \text{Inv-}\chi^2(\sigma^2|n, \nu).$$

- This can also be derived by defining the prior density for σ^2 as $p(\sigma^2) \propto 1/\sigma^2$, which is improper as

$$\int_0^\infty 1/\sigma^2 d\sigma = -\infty$$

Jeffreys' Priors

- One approach that is sometimes used to define noninformative prior distributions was introduced by Jeffreys, based on considering one-to-one transformations of the parameter: φ = h(θ).

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

- Jeffreys' general principle is that any rule for determining the prior density p(θ) should yield an equivalent result if applied to the transformed parameter
- that is, p(φ) computed by determining p(θ) should match the distribution that is obtained by determining p(φ) directly using the transformed model, p(y, φ) = p(φ)p(y|φ)

Jeffreys' Priors

- Jeffreys' principle leads to defining the noninformative prior density

$$p(\theta) = [J(\theta)]^{1/2}$$

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

$$J(\theta) = E\left[\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2 |\theta\right] = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2} |\theta\right]$$

Jeffreys' Priors

To see that Jeffreys' prior model is invariant to parameterization

$$\begin{aligned} J(\phi) &= -\mathrm{E}\left(\frac{d^2 \log p(y|\phi)}{d\phi^2}\right) \\ &= -\mathrm{E}\left(\frac{d^2 \log p(y|\theta = h^{-1}(\phi))}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) \\ &= J(\theta)\left|\frac{d\theta}{d\phi}\right|^2; \end{aligned}$$

thus,

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

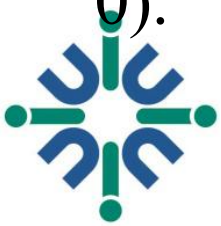
Example: binomial distribution

- log-likelihood of y ~ Bin(n, θ)

$$\log p(y|\theta) = \text{constant} + y \log \theta + (n - y) \log(1 - \theta).$$

$$J(\theta) = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2} |\theta\right] = \frac{n}{\theta(1-\theta)}$$

- Jeffreys' prior density is then $p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}$ is a Beta(1/2, 1/2) density.
- the uniform prior density, which can be expressed as θ ~ Beta(1, 1)
- p(logit(θ)) ∝ constant corresponds to the improper θ ~ Beta(0, 0).



Exercise: Normal distribution

- For the normal example with unknown variance, prove the Jeffreys prior for the standard deviation σ is

$$p(\sigma) \propto \frac{1}{\sigma}$$

- Alternative descriptions under different parameterizations for the variability are

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

$$p(\log \sigma^2) \propto p(\log \sigma) \propto 1$$

Pivotal Quantities: location parameter

- If the density of y is such that $p(y-\theta|\theta)$ is a function that is free of θ and y, say f(u), where $u = y - \theta$, then $y - \theta$ is a **pivotal quantity**, and θ is called a pure **location parameter**.

- A noninformative prior distribution should lead to posterior distribution $p(y-\theta|y)$ still be free of θ and y, $f(y-\theta)$. $y - \theta$ should still be a pivotal quantity.

- $p(y-\theta|y) \propto p(\theta)p(y-\theta|\theta)$ implying $p(\theta) \propto$ constant, a uniform distribution on θ

Pivotal Quantities: scale parameter

- If the density of y is such that p(y/θ |θ) is a function that is free of θ and y—say, g(u), where u = y/θ—then u = y/θ is a **pivotal quantity** and θ is called a pure **scale parameter**
- A noninformative prior distribution should lead to posterior distribution p(y/θ|y) still be free of θ and y, g(y/θ).
- By transformation of variables

$$p(y|\theta) = \frac{1}{\theta} p(u|\theta)$$

and similarly,

$$p(\theta|y) = \frac{y}{\theta^2} p(u|y)$$

Pivotal Quantities

- Letting both $p(u|\theta)$ and $p(u|y)$ equal $g(u)$,

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

which implies $p(\theta) \propto \frac{1}{\theta}$

- Equivalently, $p(\log \theta) \propto 1$ or $p(\theta^2) \propto \frac{1}{\theta^2}$.

- The standard deviation from a normal distribution and the mean of an exponential distribution are scale parameters.

Thanks

Multiparameter Model

Anita Wang

2017 - 2018

Introduction

- Virtually every practical problem in statistics involves more than one unknown or unobservable quantity.
- The ultimate aim of a Bayesian analysis is to obtain the **marginal posterior distribution** of the particular parameters of interest.
- We first require the joint posterior distribution of all unknowns, and then we integrate this distribution over the unknowns that to obtain the desired marginal distribution.
- In many problems there is no interest in making inferences about many of the unknown parameters. Parameters of this kind are often called **nuisance parameters**.

Averaging over 'nuisance parameters'

- Suppose θ has two parts, each of which can be a vector, $\theta = (\theta 1, \theta 2)$
- suppose that we are only interested (at least for the moment) in inference for $\theta 1$, so $\theta 2$ may be considered a **'nuisance' parameter**.
- For instance, in the simple example,

$$y|\mu, \sigma^2 \sim N(\mu, \sigma^2),$$

in which both μ (='$\theta 1$') and $\sigma 2$ (='$\theta 2$') are unknown, interest commonly centers on μ.

Joint Posterior Distribution

- We seek the conditional distribution of the parameter of interest given the observed data; in this case, $p(\theta_1|y)$.
- Joint posterior density:

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

- Marginal posterior density:

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)\, d\theta_2$$

Alternatively,

$$p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)\, d\theta_2$$

Marginal Posterior Density

- $p(\theta_1|y)$ can be regarded as a mixture of the conditional posterior distributions given the nuisance parameter,θ_2, where $p(\theta_2|y)$.is a weighting function for the different possible values of θ_2.
- The weights depend on the posterior density of θ_2 and thus on a combination of evidence from data and prior model
- $p(\theta_1|y)$ can be computed by marginal and conditional simulation, first drawing θ_2 from its marginal posterior distribution and then θ_1 from its conditional posterior distribution, given the drawn value of θ_2.

Univariate Normal with a Noninformative Prior

- Consider a vector y of n independent observations from a univariate normal distribution, $N(\mu, \sigma^2)$
- Assuming prior independence of location and scale parameters, is uniform on (μ, log σ) or,

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

- The joint posterior distribution

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

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - \mu)^2])$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

The conditional posterior distribution

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance. The sufficient statistics are $\bar{y}$ and s^2

- The conditional posterior distribution, $p(\mu|\sigma^2, y)$

we simply use the result derived in lecture 2 for the mean of a normal distribution with known variance and a uniform prior distribution:

$$\mu|\sigma^2, y \sim N(y, \sigma^2/n).$$

The marginal posterior distribution

- The marginal posterior distribution, $p(\sigma^2|y)$

$$p(\sigma^2|y) \propto \int \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu$$

$$\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{\frac{2\pi\sigma^2}{n}}$$

$$\propto (\sigma^2)^{-\frac{n+1}{2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)$$

which is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2).$$

The marginal posterior distribution

- Another way to obtain $p(\sigma^2|y)$,

$$p(\sigma^2|y) = \frac{p(\mu,\sigma^2|y)}{p(\mu|\sigma^2,y)}$$

$$= \frac{\sigma^{-n-2}\exp(-\frac{1}{2\sigma^2}[(n-1)S^2 + n(\bar{y}-\mu)^2])}{\frac{n}{\sqrt{2\pi}\sigma}\exp\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\}}$$

$$\propto \sigma^{-n-1}\exp\{-\frac{1}{2\sigma^2}[(n-1)S^2 + n(\bar{y}-\mu)^2 - n(\bar{y}-\mu)^2]\}$$

$$\propto (\sigma^2)^{-(n+1)/2}\exp(-\frac{(n-1)s^2}{2\sigma^2})$$

which is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2).$$

Joint Posterior Density

- $p(\sigma^2|y)$ is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv} - \chi^2(n-1, s^2)$$

Therefore,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Note that this result agrees with the standard frequentist result on the sample variance.

- As we know before,

$$\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n)$$

The joint posterior density,

$$p(\mu, \sigma^2|y) \propto p(\mu|\sigma^2, y)p(\sigma^2|y)$$

Sampling from the joint posterior distribution

- Now that we have $p(\mu|\sigma^2, y)$ and $p(\sigma^2|y)$, inference on μ isn't difficult.
- One method is to use the Monte Carlo approach discussed earlier
  1. Sample σ_i^2 from $p(\sigma^2|y)$
  2. Sample μ_i from $p(\mu|\sigma_i^2, y)$

Then $\mu_1, \ldots, \mu_m$ is a sample from $p(\mu|y)$.

- Note that in this case, it is actually possible to derive the exact density of $p(\mu|y)$.

Marginal Posterior Distribution for μ

- The marginal posterior distribution

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2$$

- Let $z = \frac{A}{2\sigma^2}$, where $A = (n-1)s^2 + n(\bar{y} - \mu)^2$

$$p(\mu|y) \propto A^{-\frac{n}{2}} \int_0^\infty z^{\frac{n-2}{2}} \exp(-z) dz$$

$$\propto [(n-1)s^2 + n(\bar{y} - \mu)^2]^{-\frac{n}{2}} \longleftarrow \text{Gamma integral}$$

$$\propto \left[1 + \frac{n(\bar{y} - \mu)^2}{(n-1)s^2}\right]^{-\frac{n}{2}}$$

$$\sim t_{n-1}(\bar{y}, s^2/n)$$

Marginal Posterior Distribution for μ

- The marginal posterior distribution

$$\mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

Therefore,

$$\frac{\mu - \bar{y}}{s/\sqrt{n}}|y \sim t_{n-1}$$

which corresponds to the standard result used for inference on a population mean

$$\frac{\bar{y} - \mu}{s/\sqrt{n}}|\mu, \sigma^2 \sim t_{n-1}$$

- The sampling distribution of the pivotal quantity $(\bar{y} - \mu)/(s/\sqrt{n})$ does not depend on the nuisance parameter σ^2, and its posterior distribution does not depend on data.

Posterior Predictive Distribution

- The posterior predictive distribution

$$p(\tilde{y}|y) = \iint p(\tilde{y}|\mu,\sigma^2,y)p(\mu,\sigma^2|y)\,d\mu d\sigma^2$$

$$= \int \left[\int p(\tilde{y}|\mu,\sigma^2,y)p(\mu|\sigma^2,y)\,d\mu\right] p(\sigma^2|y)\,d\sigma^2$$

$$= \int p(\tilde{y}|\sigma^2,y)p(\sigma^2|y\,)\,d\sigma^2$$

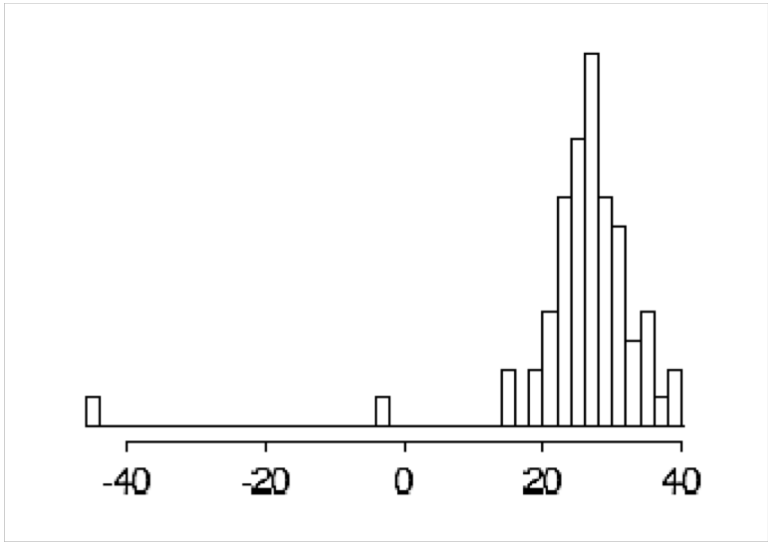
We can derive that

$$\tilde{y}|\sigma^2,y \sim N(\bar{y},\left(1+\frac{1}{n}\right)\sigma^2)$$

$$\tilde{y}|y \sim t_{n-1}(\bar{y},\left(1+\frac{1}{n}\right)s^2)$$

Example: Estimating the speed of light

- Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters. A histogram of Newcomb's 66 measurements is shown in Figure.



The data are recorded as deviations from 24,800 nanoseconds

Example

- We (inappropriately) apply the normal model, assuming that all 66 measurements are independent draws from a normal distribution with mean μ and variance σ^2. The main substantive goal is posterior inference for μ. The mean of the 66 measurements is y = 26.2, and the sample standard deviation is s = 10.8. Assuming the noninformative prior distribution $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$, a 95% central posterior interval for μ is obtained from the t_{65} marginal posterior distribution of μ as $y \pm 1.997s/\sqrt{66} = [23.6, 28.8]$.

A family of conjugate prior distributions

the conjugate prior density must also have the product form $p(\sigma^2)p(\mu|\sigma^2)$:

$$\mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{\kappa_0})$$
$$\sigma^2 \sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)$$

The joint density is

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma} \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2])$$

Conjugate Prior

- This has been labelled as $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2)$ distribution
- its four parameters can be identified as the location and scale of μ and the degrees of freedom and scale of σ^2
- One important thing to note is that with this prior, μ and σ^2 are dependent (i.e. $p(\mu|\sigma^2)$ is a function of σ^2, for example, if σ^2 is large, then a high-variance prior distribution is induced on μ
- This has a different feel from the standard frequentist analysis where $\bar{y}$ and s^2 are independent.

The Posterior Density

- The posterior density satisfies

$$p(\mu,\sigma^2|y) \propto \frac{1}{\sigma}\frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2]\right)$$

$$\times\frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\bar{y}-\mu)^2])$$

$$\propto \frac{1}{\sigma}\frac{1}{(\sigma^2)^{\frac{\nu_n}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2+\kappa_n(\mu-\mu_n)^2]\right)$$

The posterior distribution is $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_n,\frac{\sigma_n^2}{\kappa_n};\nu_n,\sigma_n^2)$

The Posterior Density

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2$$

The parameters of the posterior distribution combine the prior information and the information contained in the data. For example μ_n is a weighted average of the prior mean and the sample mean, with weights determined by the relative precision of the two pieces of information.

The Conditional Posterior Distribution $p(\mu|\sigma^2, y)$

- By using that $p(\mu|\sigma^2, y) \propto p(\mu, \sigma^2|y)$ with σ as a constant, we get

$$\mu|\sigma^2, y \sim N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

Note that the mean and variance can be written as

$$\mu_n = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} \qquad \sigma_n^2 = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}$$

which matches with the fixed variance case discuss earlier.

The Marginal Posterior Distribution $p(\sigma^2|y)$

- $p(\sigma^2|y)$

$$\sigma^2|y \sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)$$

This can be seen by the same way $p(\sigma^2|y)$ was shown in the non-informative prior case or by recognizing the $\text{N} - \text{Inv} - \chi^2$ form of the joint density.

- $p(\mu|y)$

As mentioned before, this can be determined by simulation (see in the next slide). In this case an exact answer can be determined by integrating out σ^2 from the joint density (as in the non-informative case), we get

$$\mu|y \sim t_{\nu_n}(\mu_n, \frac{\sigma_n^2}{\kappa_n})$$

Simulation of $p(\mu|y)$

- we first draw σ^2 from its marginal posterior distribution $p(\sigma^2|y)$,

$$\sigma^2|y \sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)$$

- then draw μ from its normal conditional posterior distribution $p(\mu|\sigma^2, y)$

$$\mu|\sigma^2, y \sim N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

using the simulated value of σ^2.

Multinomial Model

- The binomial distribution can be generalized to allow more than two possible outcomes.
- The multinomial sampling distribution is used to describe data for which each observation is one of k possible outcomes.
- If y is the vector of counts of the number of observations of each outcome, then

$$p(y|\theta) \propto \prod_{j=1}^{k} \theta_j^{y_j},$$

where the sum of the probabilities $\sum_{j=1}^{k} \theta_j$ is 1, and $\sum_{j=1}^{k} y_j = n$ (the number of the observations).

The Prior and Posterior Distribution

- The conjugate prior distribution

Dirichlet: a multivariate generalization of the beta distribution

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

where $\theta_j \in (0,1)$ and $\sum \theta_j = 1$

- The posterior distribution

The resulting posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$.

Prior distribution

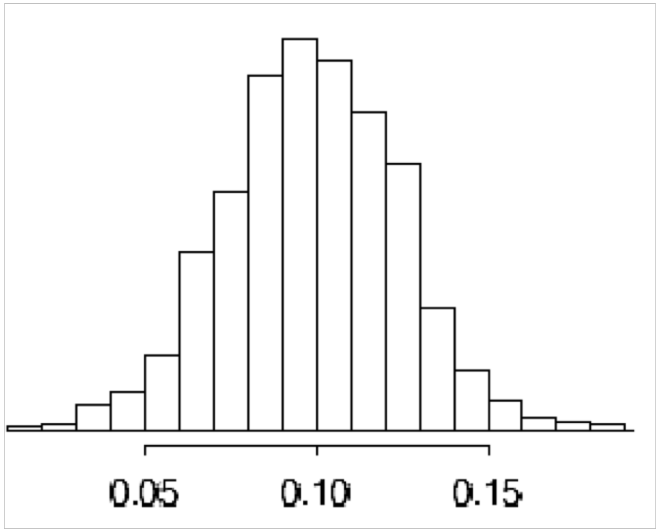
- The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{k} \alpha_j$ observations with α_j observations of the jth outcome category.
- Noninformative Dirichlet prior distributions:
  - A uniform density is obtained by setting $\alpha_j = 1$ for all j. this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^{k} \theta_j = 1$.
  - Setting $\alpha_j = 0$ for all j results in an improper prior distribution that is uniform in the $\log(\theta_j)$'s. The resulting posterior distribution is proper if there is at least one observation in each of the k categories,

Example: Pre-election polling

- In late October, 1988, a survey was conducted by CBS News of 1447 adults in the United States to find out their preferences in the upcoming presidential election. Out of 1447 persons, y_1 = 727 supported George Bush, y_2 = 583 supported Michael Dukakis, and y_3 = 137 supported other candidates or expressed no opinion.
- then the data (y_1, y_2, y_3) follow a multinomial distribution, with parameters $(\theta_1, \theta_2, \theta_3)$, the proportions of Bush supporters, Dukakis supporters, and those with no opinion in the survey population.
- An estimand of interest is $\theta_1 - \theta_2$, the population difference in support for the two major candidates

Example

- With a noninformative uniform prior distribution on θ, $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is Dirichlet(728, 584, 138).
- We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3)$ from the posterior Dirichlet distribution and then compute θ1 − θ2 for each.



All of the 1000 simulations had $\theta_1 > \theta_2$; thus, the estimated posterior probability that Bush had more support than Dukakis in the survey population is over 99.9%.

Multivariate Normal Model

- y is a vector of length d with mean vector μ (also of length d and $d \times d$ variance matrix Σ), with multivariate normal distribution

$$y|\mu, \Sigma \sim N_d(\mu, \Sigma)$$

- The density of a single observation is

$$p(y|\mu, \Sigma) \propto |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right)$$

- The likelihood of n i.i.d observations is

$$p(y_1, \ldots, y_n|\mu, \Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T \Sigma^{-1}(y_i-\mu)\right)$$

$$= |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1} S_0)\right)$$

Multivariate Normal Model

where $\text{tr}(A)$ is the trace of the matrix A (the sum of the diagonal entries) and

$$S_0 = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T$$

- So the density and likelihood look like what we get in the univariate case, but with matrix and vectors instead.
- Note that most of the inference in this model is a direct analogue to the univariate case. However we need a multivariate analogue to the χ^2 and Inv-χ^2 distributions.

Wishart and Inverse Wishart Distributions

- Wishart distribution ($\text{Wishart}_\nu(\Lambda)$)
- Multivariate analogue of a scaled χ^2 distribution

If $z_1, \ldots, z_\nu \sim N_d(0, \Lambda)$ then

$$\Sigma = \sum_{i=1}^{\nu} z_i z_i^T \sim \text{Wishart}_\nu(\Lambda)$$

like $z_1, \ldots, z_\nu \sim N_d(0, \tau^2)$ then

$$S = \sum_{i=1}^{\nu} z_i^2 \sim \tau^2 \chi_\nu^2$$

Inverse Wishart Distribution

- Inverse Wishart distribution ($\text{Inv} - \text{Wishart}_\nu(\Lambda^{-1})$)
- Multivariate analogue of a scaled $\text{Inv} - \chi^2$ distribution

If $\Sigma \sim \text{Wishart}_\nu(\Lambda)$ then

$$\Sigma^{-1} \sim \text{Inv} - \text{Wishart}_\nu(\Lambda^{-1})$$

Multivariate Normal Models

- Unknown mean μ but known Σ

The conjugate prior distribution for μ is

$$\mu|\Sigma \sim N(\mu_0, \Lambda_0)$$

The posterior density

$$\mu|\Sigma, y \sim N(\mu_n, \Lambda_n)$$

where

$$\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$
$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

Like the univariate case, the posterior mean is a weighted average of the prior mean and the sample average and the posterior precision matrix is the prior precision matrix + data precision matrix.

The Conditional Posterior Distribution

- The marginal posterior distribution of a subset of the parameters, $\mu^{(1)}$ say, is also multivariate normal, with mean vector equal to the appropriate subvector of the posterior mean vector μ_n and variance matrix equal to the appropriate submatrix of Λ_n.
- The conditional posterior distribution of a subset $\mu^{(1)}$ given the values of a second subset $\mu^{(2)}$ is multivariate normal.

$$\mu^{(1)}|\mu^{(2)}, y \sim N\left(\mu_n^{(1)} + \beta^{1|2}\left(\mu^{(2)} - \mu_n^{(2)}\right), \Lambda^{1|2}\right),$$

where the regression coefficients $\beta^{1|2}$ and conditional variance matrix $\Lambda^{1|2}$ are defined by

$$\beta^{1|2} = \Lambda_n^{(12)}\left(\Lambda_n^{(22)}\right)^{-1}$$

$$\Lambda^{1|2} = \Lambda_n^{(11)} - \Lambda_n^{(12)}\left(\Lambda_n^{(22)}\right)^{-1}\Lambda_n^{(21)}$$

Unknown Mean and Variance

- Conjugate Prior

$$\Sigma \sim \text{Inv} - \text{Wishart}_{\nu_0}(\Lambda_0^{-1})$$
$$\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$$

- The posterior distribution satisfies

$$\Sigma|y \sim \text{Inv} - \text{Wishart}_{\nu_n}(\Lambda_n^{-1})$$
$$\mu|\Sigma, \text{y} \sim \text{N}(\mu_n, \Sigma/\kappa_n)$$

The Posterior Distribution

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$$

$$S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$$

Marginal Distribution

- In addition, it is possible to integrate out the variance matrix showing that

$$\mu|y \sim \mathrm{t}_{\nu_n - \mathrm{d}+1}(\mu_n, \Lambda_n/(\kappa_n(\nu_n - d + 1)))$$

(i.e. multivariate t with $\nu_n - d + 1$ degrees of freedom)

Thanks

Hierarchical models

Anita Wang

2017 - 2018

Constructing a parameterized prior distribution

- we consider the problem of estimating a parameter θ using data from a small experiment and a prior distribution constructed from similar previous (or historical) experiments.
- Mathematically, we will consider the current and historical experiments to be a random sample from a common population.

Example

- In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents
- Suppose the immediate aim is to estimate θ, the probability of tumor in a population of female laboratory rats that receive a zero dose of the drug
- The data show that 4 out of 14 rats developed tumor
- It is natural to assume a binomial model for the number of tumors, given θ.
- For convenience, we select a prior distribution for θ from the conjugate family, $\theta \sim \text{Beta}(\alpha, \beta)$.

Analysis with a fixed prior distribution

- From historical data, suppose we knew that the tumor probabilities θ among groups of female lab rats follow an approximate beta distribution, with known mean and standard deviation.
- we could find values for α, β that correspond to the given values for the mean and standard deviation.
- assuming a Beta(α, β) prior distribution for θ yields a Beta(α + 4, β + 10) posterior distribution for θ.

Approximate estimate of the population distribution

- Typically, the mean and standard deviation of underlying tumor risks are not available.
- historical data are available on previous experiments on similar groups of rats.
- the historical data were in fact a set of observations of tumor incidence in 70 groups of rats. In the jth historical experiment, let the number of rats with tumors be y_j and the total number of rats be n_j.
- We model the y_j's as independent binomial data, given sample sizes n_j and study-specific means θ_j. Assuming that the beta prior distribution with parameters (α, β) is a good description
- of the population distribution of the θ_j 's in the historical experiments

Table: tumor incidence

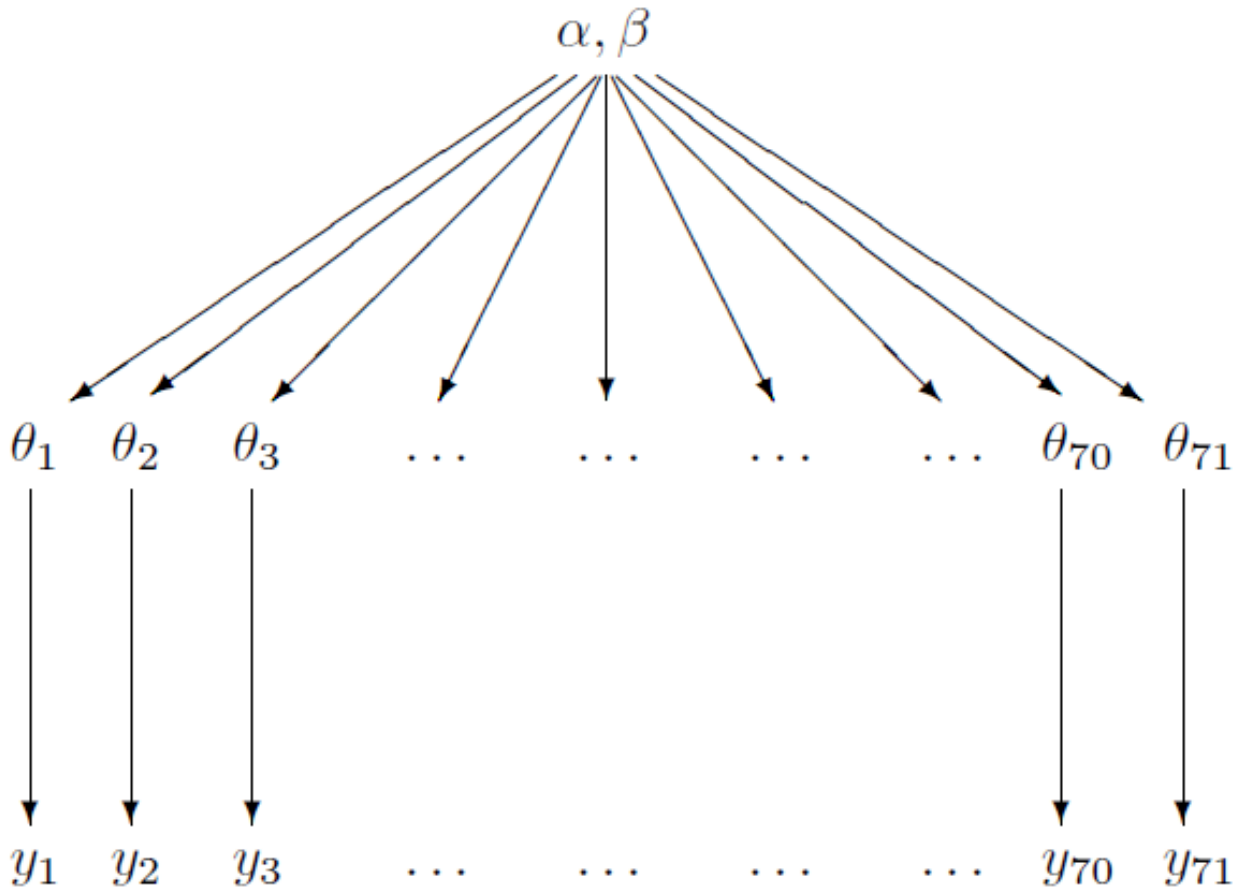
Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:

4/14

Structure of the hierarchical model



A starting point

- The observed sample mean and standard deviation of the 70 values $\frac{y_j}{n_j}$ are 0.136 and 0.103.
- If we set the mean and standard deviation of the population distribution to these values, we can solve for α and β. The resulting estimate for (α, β) is (1.4, 8.6).
- This is not a Bayesian calculation because it is not based on any specified full probability model. The estimate (1.4, 8.6) is simply a starting point from which we can explore the idea of estimating the parameters of the population distribution.

Exchangeability

- Generalizing from the example of the previous section, consider a set of experiments j = 1, . . . , J, in which experiment j has data (vector) y_j and parameter (vector) θ_j , with likelihood p(y_j |θ_j).
- A useful assumption in building models, if no information, other than the data y is available to distinguish any of the μ_j's from any of the others, and no ordering of grouping of the parameters can be made, one must assume symmetry among the parameters in the prior.

Exchangeability

- For example, in the rat tumor example, we have no prior reason to assume that $\mu_{70} < \mu_{71}$ is more likely than $\mu_{70} > \mu_{71}$. In fact, for the information given, the order that the groups are listed in is meaningless.
- So for this problem, it seems reasonable to have the distribution on the μ_j's be exchangeable, i.e. the distribution $p(\mu_1, \ldots \mu_J)$ should be invariant under permutations of the indices $(1, \ldots, J)$. If J = 3, then the distributions

$p(\theta_1, \theta_2, \theta_3), p(\theta_1, \theta_3, \theta_2), p(\theta_2, \theta_1, \theta_3), p(\theta_2, \theta_3, \theta_1),$
$p(\theta_3, \theta_1, \theta_2), p(\theta_3, \theta_2, \theta_1)$

are all of the same form.

Objections to exchangeable models

- it is natural to object to exchangeability on the grounds that the units actually differ.
- For example, the 71 rat tumor experiments were performed at different times, on different rats, and presumably in different laboratories. Such information does not, however, invalidate exchangeability
- Note that exchangeability does not imply independence.
- However all iid models are exchangeable.

Exchangeability

- One way of getting exchangeable distribution is to take a mixture of iid distributions.

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

- As ϕ is usually unknown, the distribution on μ must average over the uncertainty in ϕ.

$$p(\theta) = \int \left[\prod_{j=1}^{J} p(\theta_j|\phi) \right] p(\phi) d\phi$$

- All models of this form are exchangeable. To think of the μ_j's as draws from a superpopulation model that is determined by the hyperparameter ϕ.

Exchangeability

- One way of thinking of exchangeability is in terms non-informativeness or ignorance about the random variables.
- In the rat example, we have no preferences for different orderings of the theta's.
- In the rat problem, for example, the model like

$$\theta_i \sim Beta(\alpha, \beta) \text{ iid}$$
$$\alpha \sim U(0, 20)$$
$$\beta \sim U(0, 20)$$

Bayesian treatment of the hierarchical model

- the key 'hierarchical' part of these models is that φ is not known and thus has its own prior distribution, p(φ).
- Suppose we have the following hierarchical model

$$y|\theta,\phi \sim p(y|\theta)$$
$$\theta|\phi \sim p(\theta|\phi)$$
$$\phi \sim p(\phi)$$

- The joint prior distribution is

$$p(\varphi, \theta) = p(\varphi)p(\theta|\varphi),$$

and the joint posterior distribution is

$$p(\varphi, \theta|y) \propto p(\varphi, \theta)p(y|\varphi, \theta)$$
$$= p(\varphi, \theta)p(y|\theta),$$

The hyperprior distribution

- In order to create a joint probability distribution for (φ, θ), we must assign a prior distribution to φ.
- If little is known about φ, we can assign a diffuse prior distribution, but we must be careful when using an improper prior density to check that the resulting posterior distribution is proper
- In most real problems, one should have enough substantive knowledge about the parameters in φ
- In the rat tumor example, the hyperparameters are (α, β), which determine the beta distribution for θ.

Posterior predictive distributions

- There are two posterior predictive distributions that might be of interest to the data analyst:
  - the distribution of future observations $\tilde{y}$ corresponding to an existing θ_j
  - the distribution of observations $\tilde{y}$ corresponding to future θ_j 's ($\tilde{\theta}$) drawn from the same superpopulation.
- In the rat tumor example, future observations can be (1) additional rats from an existing experiment, or (2) results from a future experiment.

Bayesian analysis of conjugate hierarchical models

- we present an approach that combines analytical and numerical methods to obtain simulations from the joint posterior distribution, $p(\theta, \varphi|y)$
- the population distribution, $p(\theta|\varphi)$, is conjugate to the likelihood, $p(y|\theta)$.
- Three-step:
  1. $p(\varphi, \theta|y) \propto p(\varphi)p(\theta|\varphi)p(y|\theta, \varphi)$
  2. Conditional posterior
     $p(\theta|\varphi,y) \propto p(\theta|\varphi)p(y|\theta, \varphi)$
     This will be easy if a conjugate prior is used.
  3. Marginal posterior
     $p(\varphi|y) = \int p(\theta, \varphi|y)d\theta$. .

Bayesian analysis of conjugate hierarchical models

- For many standard models, the marginal posterior distribution of φ can be computed using the conditional probability,

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

- The difficulty is that the denominator $p(\theta|\varphi, y)$, a function of both θ and φ for fixed y, has a normalizing factor that depends on φ as well as y.
- One must be careful with the proportionality 'constant' in Bayes' theorem, especially when using hierarchical models, to make sure it is actually constant

Drawing simulations from the posterior distribution

- Simulating a draw from the joint posterior distribution, p(θ, φ|y):
  1. Sample $\phi_1, \ldots, \phi_m$ from p(φ|y)
  2. Sample $\theta_1, \ldots, \theta_m$ from p(θ|φ_i, y), given the drawn value of φ.
  3. If necessary, sample $\tilde{y}$. The form of this draw depends on whether the θ of interest is one corresponding to the dataset or a new one. It might be necessary first to draw a new value $\tilde{\theta}$ given φ.

  As usual, the above steps are performed L times in order to obtain a set of L draws.

Application to the model for rat tumors

- the data from experiments j = 1, . . . , J, J = 71, are assumed to follow independent binomial distributions:

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

- The parameters θ_j are assumed to be independent samples from a beta distribution:

$$\theta_j \sim \text{Beta}(\alpha, \beta),$$

- we shall assign a noninformative hyperprior distribution to reflect our ignorance about the unknown hyperparameters.

Joint, conditional, and marginal posterior distributions

- The joint posterior distribution of all parameters is

$$p(\theta, \alpha, \beta|y) \propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta)$$

$$\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1}(1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j}(1-\theta_j)^{n_j-y_j}$$

- Given (α, β), the components of θ have beta densities, and the joint density is

$$p(\theta|\alpha, \beta, y)$$

$$= \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}$$

Joint, conditional, and marginal posterior distributions

- Marginal posterior

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

$$\propto p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma\left(\alpha+y_j\right)\Gamma\left(\beta+n_j-y_j\right)}{\Gamma\left(\alpha+\beta+n_j\right)}$$

‘noninformative’ hyperprior distribution

- One reasonable choice of diffuse hyperprior density is uniform on $(\alpha/\alpha+\beta\ ,\ (\alpha+\beta)^{-1/2})$,

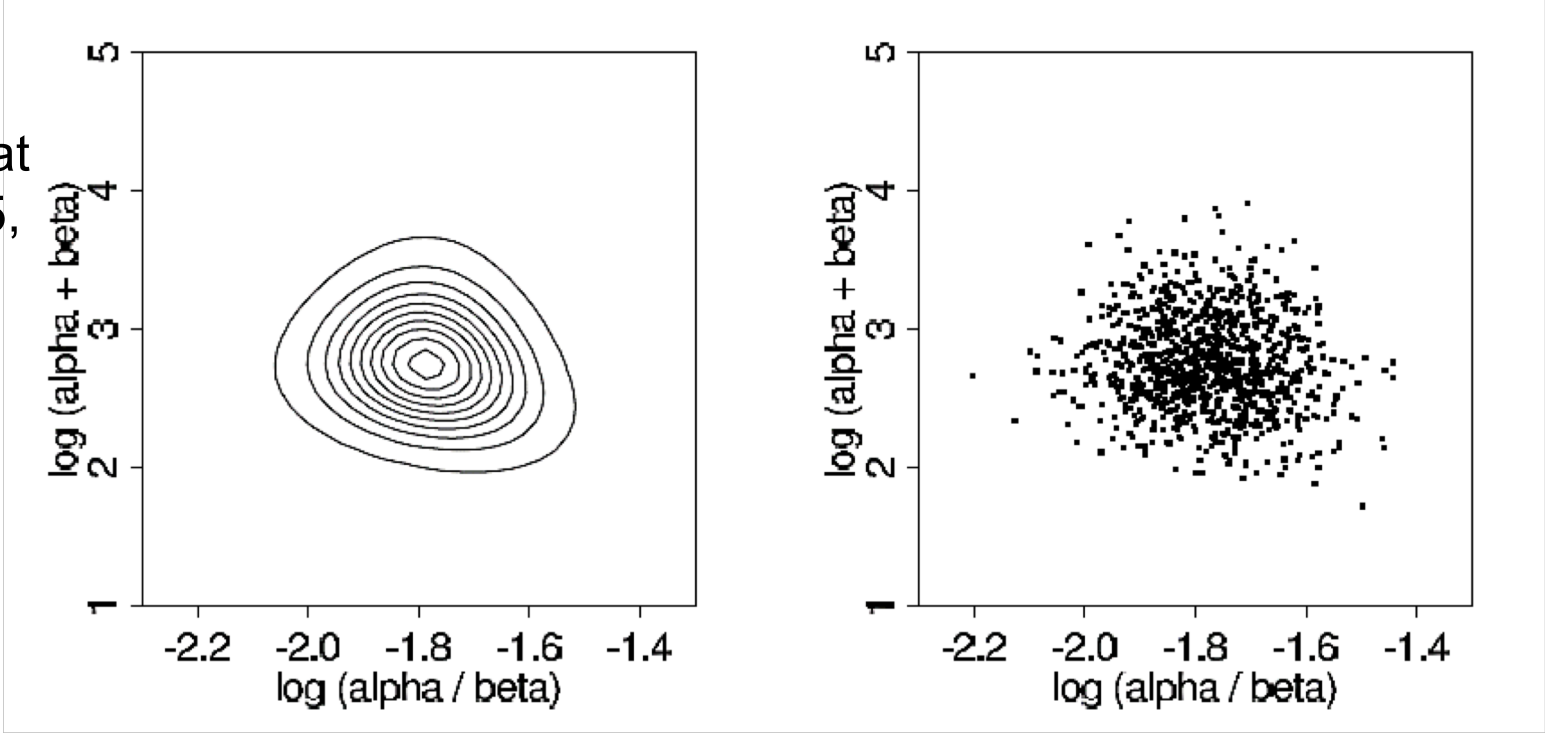
$$p(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$$

- Before assigning a hyperprior distribution, we reparameterize in terms of $\text{logit}(\alpha/\alpha+\beta\ ) = \log(\alpha/\beta\ )$ and $\log(\alpha+\beta)$, which are the logit of the mean and the logarithm of the ‘sample size’ in the beta population distribution for θ.
- We use the logistic and logarithmic transformations to put each on a $(-\infty,\infty)$ scale.

the marginal posterior density of the hyperparameters

- the mode ((-1.75, 2.8) and $(\alpha, \beta) = (2.4, 14.0)$) is not far from the point estimate (as we would expect)
- important parts of the marginal posterior distribution lie outside the range of the graph.

Contour lines are at 0.05, 0.15, . . . , 0.95 times the density at the mode



posterior moments

- E(α|y) is estimated by

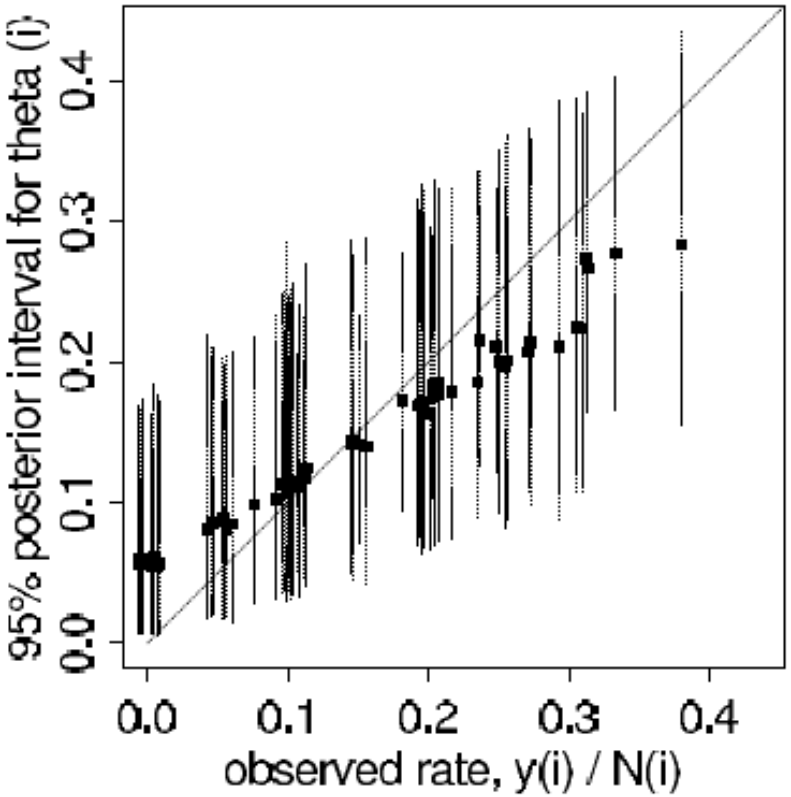
$$\sum_{\log\left(\frac{\alpha}{\beta}\right),\log(\alpha+\beta)} \alpha p\left(\log\left(\frac{\alpha}{\beta}\right), \log(\alpha+\beta)\,|y\right)$$

- we compute E(α|y) = 2.4 and E(β|y) = 14.3.

Sampling from the joint posterior distribution

- We draw 1000 random samples from the joint posterior distribution of $(\alpha, \beta, \theta_1, \ldots, \theta_J)$
  1. Simulate 1000 draws of $(\log(\alpha/\beta), \log(\alpha+\beta))$ from their posterior distribution displayed in Figure
  2. For $l = 1, \ldots, 1000$:
     a) Transform the lth draw of $(\log(\alpha/\beta), \log(\alpha+\beta))$ to the scale (α, β) to yield a draw of the hyperparameters from their marginal posterior distribution
     b) For each $j = 1, \ldots, J$, sample θ_j from its conditional posterior distribution, $\theta_j|\alpha, \beta, y \sim \text{Beta}(\alpha + y_j, \beta + n_j - y_j)$.

Displaying the results



Posterior medians and 95% intervals of rat tumor rates, θ_j, (plotted vs. observed tumor rates y_j/n_j), based on simulations from the joint posterior distribution. The 45◦ line corresponds to the unpooled estimates, $\hat{\theta}= y_i/n_i$.

Thanks

Model Checking

Anita Wang

2017 - 2018

Model Checking

"All models are wrong but some models are useful"

--- George E. P. Box

- Checking the model is crucial to statistical analysis.
- Bayesian prior-to-posterior inferences assume the whole structure of a probability model and can yield misleading inferences when the model is poor.
- A good Bayesian analysis, therefore, should include at least some check of the adequacy of the fit of the model

Model Checking

- So far we have looked at a number of models and examined them with example data sets. Do the models used accurately describe the data used?
- In standard analyses, we will often check model assumptions. For example, in standard regression we will check for
  - Correct form of the regression function (e.g. linear vs quadratic)
  - Constant variance of the residuals
  - Independence and normality of the residuals

Model Checking

- Basic question: How sensitive are our posterior inferences to our modelling assumptions?
- Student sleeping time: Will the following models give significantly different answers about the probability of heavy sleepers?
- Original Model:
  - Data model: y is the number of heavy sleepers

$$y|\theta \sim Bin(n, \theta)$$

  - Prior: θ the probability of heavy sleepers

$$\theta \sim Beta(\alpha, \beta)$$

Model Checking

- Alternative model 1:
  - Data model: y is the number of heavy sleepers

$$y|\theta \sim Bin(n, \theta)$$

  - Prior: θ the probability of heavy sleepers

$$logit(\theta) \sim N(\mu, \sigma^2)$$

where $logit(\theta) = \log\frac{\theta}{1-\theta}$

- Alternative model 2:
  - Data model: y is the number of heavy sleepers

$$y|\theta \sim Beta - bin(n, \alpha, \beta)$$

  - Prior: (α, β) probability parameters

$$\alpha, \beta \sim Gamma(\gamma_\alpha, \delta_\alpha)Gamma(\gamma_\beta, \delta_\beta)$$

Model Checking

- Note that we will not be trying to answer the question of whether our model is correct or not. We are interested in whether the inaccuracies matter.
- One approach to build a super-model that contains all of our models of interest as special cases. This approach usually isn't taken as it is usually difficult to build this super-model and computation is usually infeasible, assuming you can build the model.
- Instead we will base these checks on the posterior predictive distribution. Does our data look like our fitted model says it should.

Model Checking

- Check on the posterior predictive distribution:
  - *External validation*: future data is compared with the posterior predictive distribution.
  - *Internal validation*: observed data is compared with the posterior predictive distribution.

- Compare the outliers and graphs of the future (existing) data and simulated data from posterior predictive distribution $p(\tilde{y}|y)$

Example: speed of light

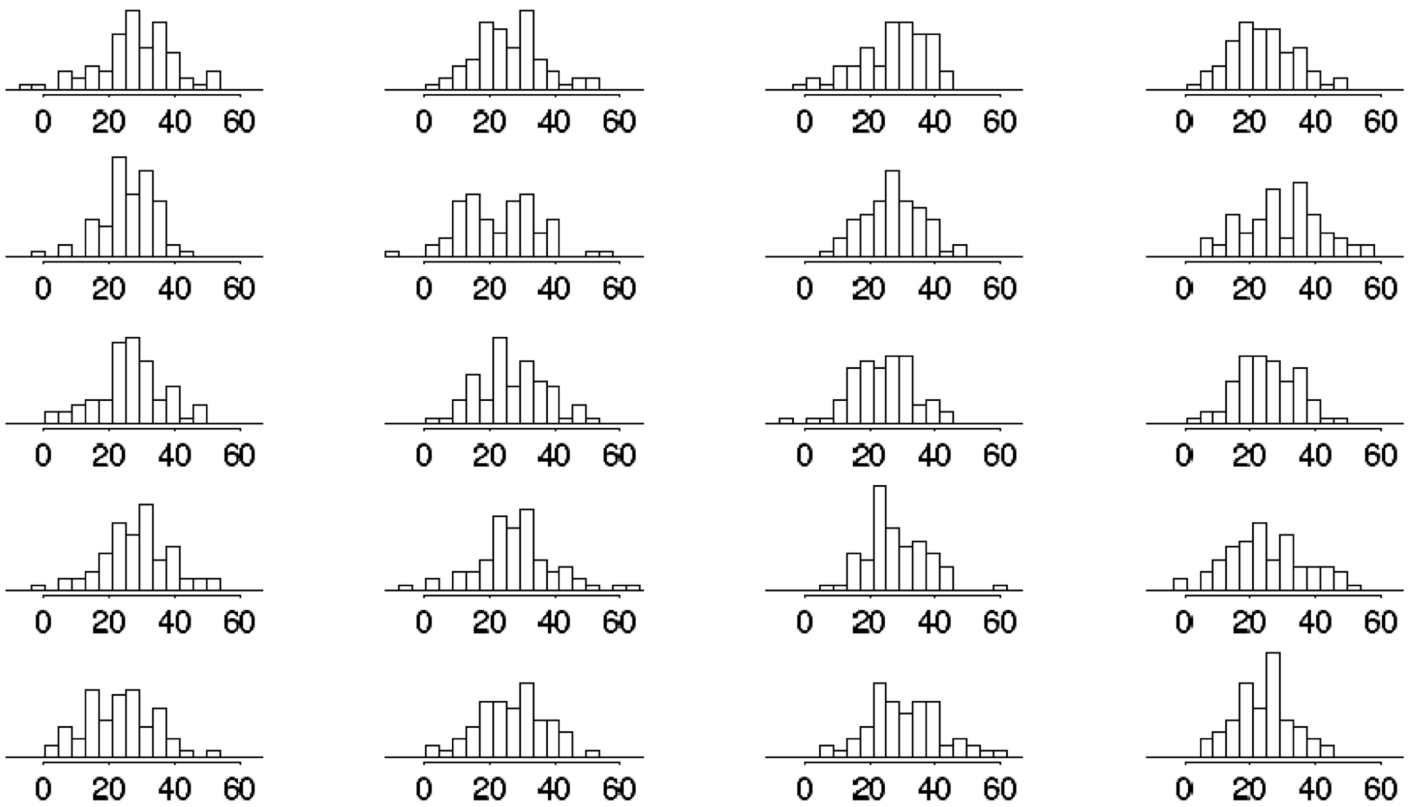


Figure 6.2 *Twenty replications, y^{rep}, of the speed of light data from the posterior predictive distribution, $p(y^{\text{rep}}|y)$; compare to observed data, y, in Figure 3.1. Each histogram displays the result of drawing 66 independent values $\tilde{y}_i$ from a common normal distribution with mean and variance (μ, σ^2) drawn from the posterior distribution, $p(\mu, \sigma^2|y)$, under the normal model.*

Example: speed of light

- Figure displays twenty histograms, each of which represents a single draw from the posterior predictive distribution of the values in Newcomb's experiment
- first drawing (μ, σ^2) from their joint posterior distribution, then drawing 66 values from a normal distribution with this mean and variance.
- All these histograms look different from the histogram of actual data in the next slide
- One way to measure the discrepancy is to compare the smallest value in each hypothetical replicated dataset to Newcomb's smallest observation, −44.

Example: speed of light

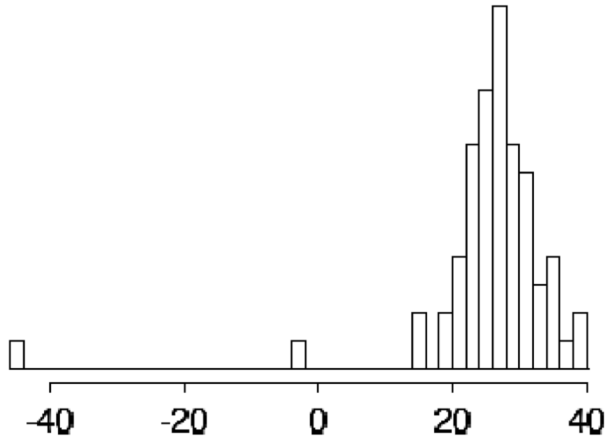


Figure 3.1 *Histogram of Simon Newcomb's measurements for estimating the speed of light, from Stigler (1977). The data are recorded as deviations from* 24,800 *nanoseconds.*

Example: speed of light

The histogram below shows the smallest observation in each of the 20 hypothetical replications; all are much larger than Newcomb's smallest observation

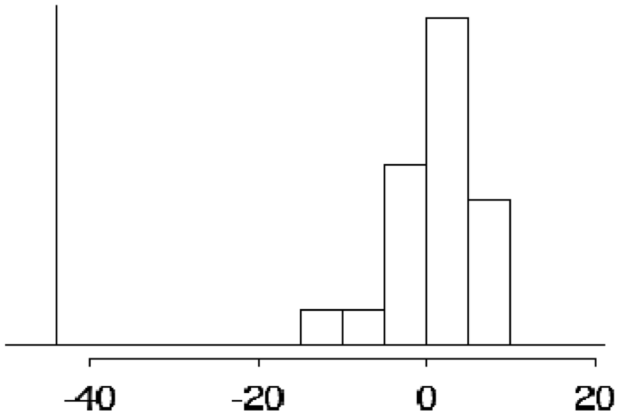
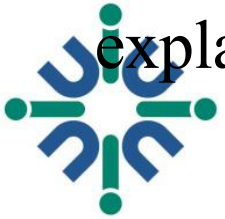


Figure 6.3 *Smallest observation of Newcomb's speed of light data (the vertical line at the left of the graph), compared to the smallest observations from each of the 20 posterior predictive simulated datasets displayed in Figure 6.2.*

Posterior Predictive Checking

- Idea: If the model fits, replicated data generated under the model should look similar to the observed data.
- If we see some discrepancy, is it due to model misspecification or due to chance.
- Approach: Generate L datasets, $y_1^{rep}, \ldots, y_L^{rep}$ from the posterior predictive distribution $p(y^{rep}|y)$. y^{rep} corresponds to replicated data.
- $\tilde{y}$ represents any future outcome whereas y^{rep} indicates a replication exactly like the observed y.
- For example, if the model has explanatory variables, x, they will be identical for y and y^{rep}, but $\tilde{y}$ may have its own explanatory variables, $\tilde{x}$.



Test quantities

- We measure the discrepancy between model and data by defining test quantities
- A test quantity (discrepancy measure), $T(y, \theta)$, is a scalar summary of parameters and data when comparing data to predictive simulations.
- We use the notation $T(y)$ for a test statistic, which is a test quantity that depends only on data
- In the Bayesian context, we can generalize test statistics to allow dependence on the model parameters under their posterior distribution

Tail-area probabilities

- The lack of fit of the data as compared to the posterior predictive distribution can be compared by a tail-area probability (e.g. p-value) of the test statistic $T(y, \theta)$. To calculate this probability we will use the replicates sampled from $p(y^{rep}|y)$.
- Classical p-value

$$p_C = P[T(y^{rep}) \geq T(y)|\theta]$$

where the probability is calculated over the distribution of y^{rep} given a fixed θ. In the classical testing setting θ would correspond to the null hypothesis value. It could also be a point estimate (say the MLE).

Posterior predictive *p*-values

- To evaluate the fit of the posterior distribution of a Bayesian model, we can compare the observed data to the posterior predictive distribution
- In the Bayesian approach, test quantities can be functions of the unknown parameters as well as data
- The Bayesian p-value is defined as the probability that the replicated data could be more extreme than the observed data,

$$p_B = P[T(y^{rep}, \theta) \geq T(y, \theta)|y]$$

$$= \iint I\big(T(y^{rep}, \theta) \geq T(y, \theta)\big) p(y^{rep}|\theta) p(\theta|y)\, dy^{rep} d\theta$$

Posterior predictive *p*-values

- Usually we can't calculate the Bayesian p-value exactly, but can do it by simulation. Suppose that we have L simulations of $\theta(\theta^1, \ldots, \theta^L)$ from the posterior distribution $p(\theta|y)$. Then for each of these θ samples, generate one sample y^{rep_l} from $p(y^{rep}|\theta^l)$.
- We want to compare each of the $T(y^{repl}, \theta^l)$ with $\mathrm{T}(y, \theta^l)$ Then

$$\hat{p}_B = \frac{1}{L}\sum_{l=1}^{L} I(T(y^{repl}, \theta^l) \geq \mathrm{T}(y, \theta^l))$$

(i.e. the proportion of samples where $T(y^{repl}, \theta^l) \geq \mathrm{T}(y, \theta^l)$) is an estimate of p_B.

Posterior predictive *p*-values

- Note that the test statistic $T(y, \theta)$ needs to be chosen to investigate deviations of interest. This is similar to choosing a powerful test statistic when conducting a hypothesis test
- For example, in the analysis of Newcomb's speed of light experiment, a worry was the effect of outliers. Thus $T(y, \theta)$ needs to be chosen to focus on this issue.
- Previously, we use $T(y,\theta) = \min y_i$ as test statistics to demonstrate the poor fit of the normal model to the speed of light data
- We try to use other test quantities

Posterior predictive *p*-values

- Variance:

The sample variance does not make a good test statistic because it is a sufficient statistic of the model

With non-informative prior, the posterior distribution will automatically be centered near the observed value.

- A test quantity sensitive to asymmetry

$$T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|.$$

The 61st and 6th order statistics are chosen to represent approximately the 90% and 10% points of the distribution. The test quantity should be scattered about zero for a symmetric distribution.

Choosing test quantities

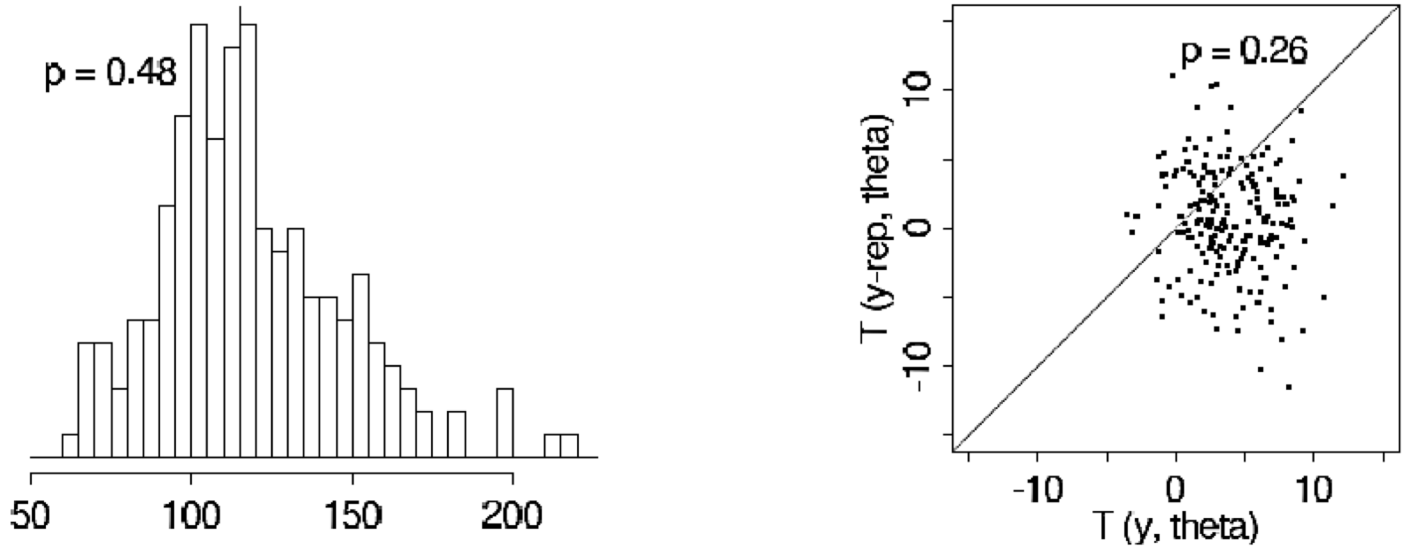
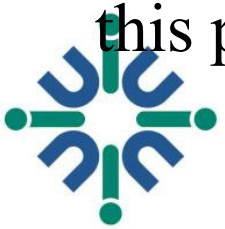


Figure 6.4 *Realized vs. posterior predictive distributions for two more test quantities in the speed of light example: (a) Sample variance (vertical line at 115.5), compared to 200 simulations from the posterior predictive distribution of the sample variance. (b) Scatterplot showing prior and posterior simulations of a test quantity:* $T(y,\theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$ *(horizontal axis) vs.* $T(y^{\text{rep}},\theta) = |y^{\text{rep}}_{(61)} - \theta| - |y^{\text{rep}}_{(6)} - \theta|$ *(vertical axis) based on 200 simulations from the posterior distribution of* (θ, y^{rep})*. The p-value is computed as the proportion of points in the upper-left half of the scatterplot.*

Interpreting posterior predictive p-values

- Major failures of the model, typically corresponding to extreme tail-area probabilities (less than 0.01 or more than 0.99.
- Lesser failures might also suggest model improvements or might be ignored in the short term if the failure appears not to affect the main inferences.
- even extreme p-values may be ignored if the misfit of the model is substantively small compared to variation within the model.
- We typically evaluate a model with respect to several test quantities, and we should be sensitive to the implications of this practice.



Graphical posterior predictive checks

- Direct display of all the data
- Display of data summaries or parameter inferences.

This can be useful in settings where the dataset is large and we wish to focus on the fit of a particular aspect of the model.

- Graphs of residuals or other measures of discrepancy between model and data.

Direct data display

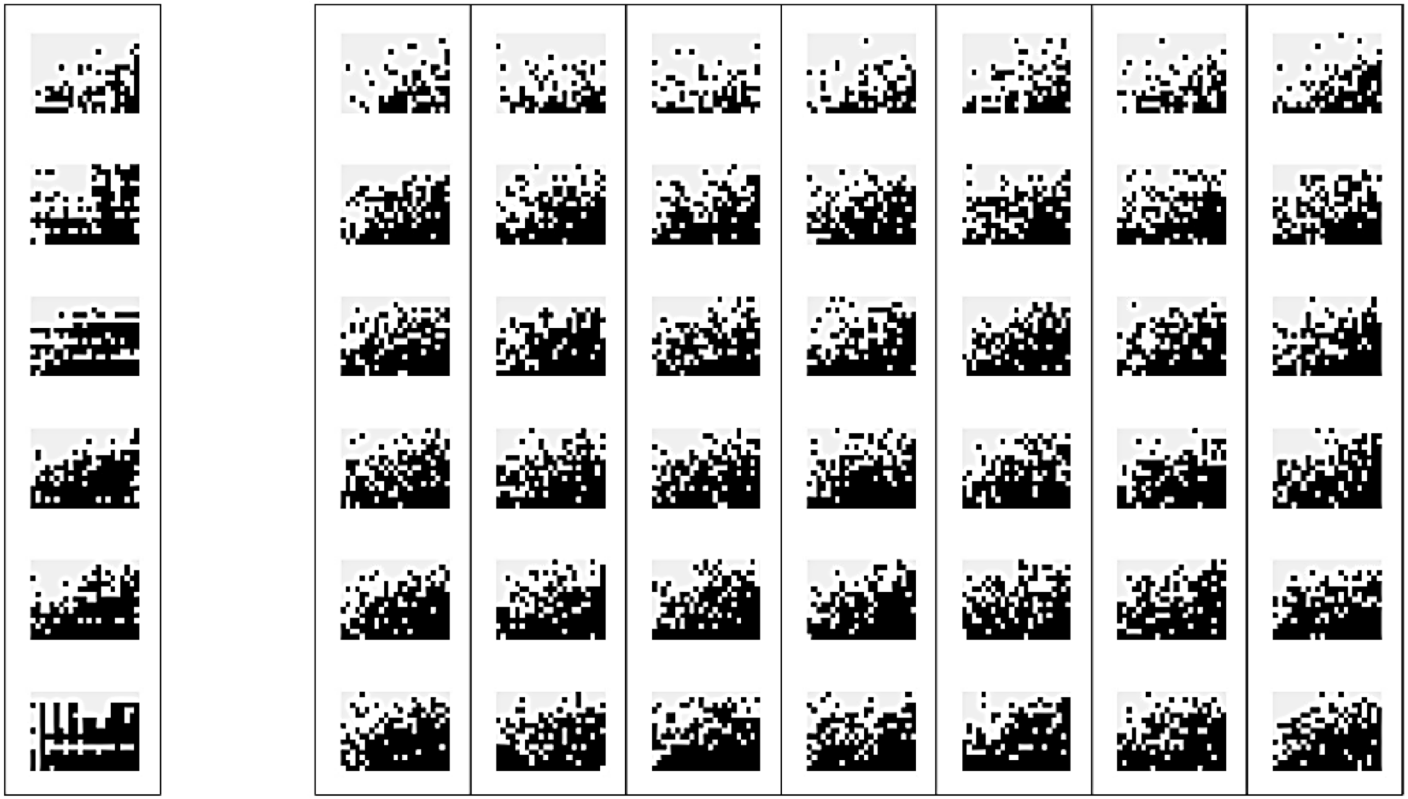


Figure 6.7 *Left column displays observed data y (a 15 × 23 array of binary responses from each of 6 persons); right columns display seven replicated datasets y^{rep} from a fitted logistic regression model. A misfit of model to data is apparent: the data show strong row and column patterns for individual persons (for example, the nearly white row near the middle of the last person's data) that do not appear in the replicates. (To make such patterns clearer, the indexes of the observed and each replicated dataset have been arranged in increasing order of average response.)*

Direct data display

- The left column of the figure displays binary data

for each of 6 persons, a possible ‘yes’ or ‘no’ to each of 15 possible reactions (displayed as rows) to 23 situations (columns)—from an experiment in psychology

- The right columns of the figure display seven independently simulated replications y^{res} from a fitted logistic regression model
- Before displaying, the reactions, situations, and persons have been ordered in increasing average response.
- the replicated datasets look ‘random’ compared to the observed data, which have strong rectilinear structures that are clearly not captured in the model.

Displaying summary statistics or inferences

- the model included two vectors of parameters, $\varphi_1, \ldots, \varphi_{90}$, and $\psi_1, \ldots, \psi_{69}$, corresponding to patients and psychological symptoms, and that each of these 159 parameters were assigned independent Beta(2, 2) prior distributions.
- Data were collected (measurements of which symptoms appeared in which patients) and the full Bayesian model was fitted, yielding posterior simulations for all these parameters.
- If the model were true, we would expect any single simulation draw of the vectors of patient parameters φ and symptom parameters ψ to look like independent draws from the Beta(2, 2) distribution.

Displaying summary statistics or inferences

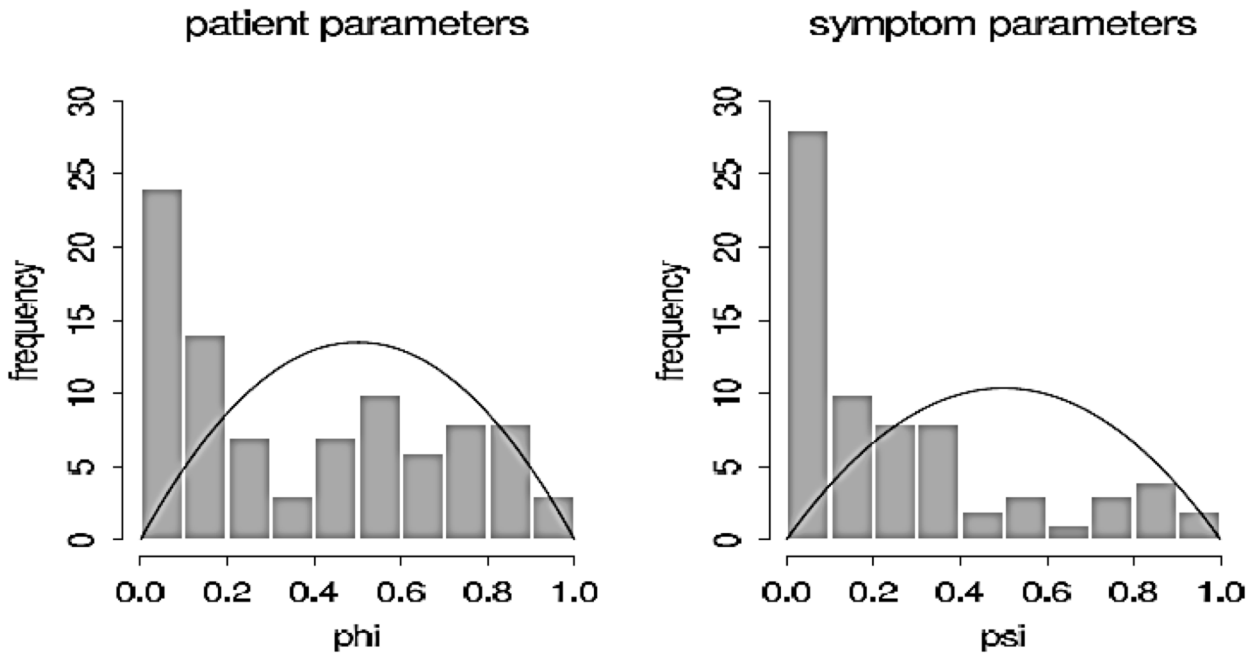


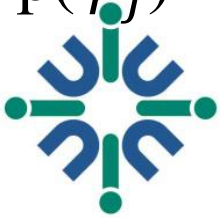
Figure 6.9 *Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, from a single draw from the posterior distribution of a psychometric model. These histograms of posterior estimates contradict the assumed Beta(2, 2) prior densities (overlain on the histograms) for each batch of parameters, and motivated us to switch to mixture prior distributions. This implicit comparison to the values under the prior distribution can be viewed as a posterior predictive check in which a new set of patients and a new set of symptoms are simulated.*

Displaying summary statistics or inferences

- as a model check we can plot a histogram of a single simulation of the vector of parameters φ or ψ and compare to the prior distribution
- The lines in the figure show the Beta(2, 2) prior distribution, which clearly does not fit.
- For both φ and ψ, there are too many cases near zero, corresponding to patients and symptoms that almost certainly are not associated with a particular syndrome.
- Consider:

$$p(\varphi_j) = 0.5\ \text{Beta}(\varphi_j|1, 6) + 0.5\ \text{Beta}(\varphi_j|1, 1)$$

$$p(\psi_j) = 0.5\ \text{Beta}(\psi_j|1, 16) + 0.5\ \text{Beta}(\psi_j|1, 1).$$



Displaying summary statistics or inferences

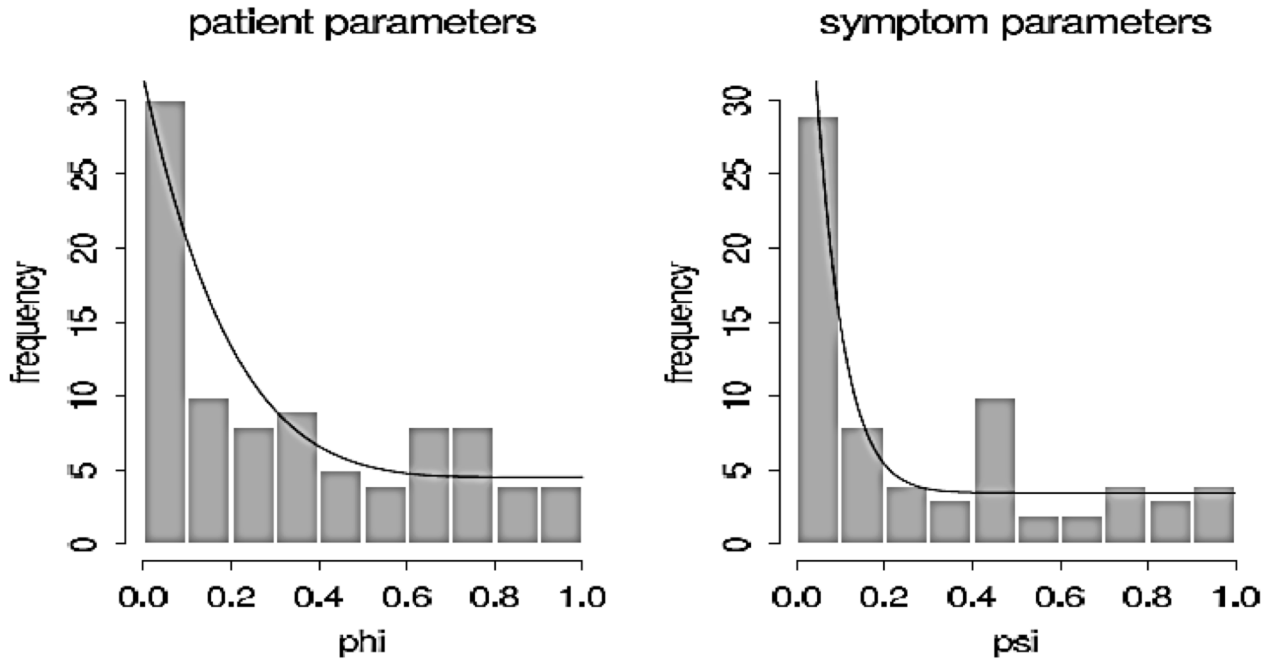


Figure 6.10 *Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, as estimated from an expanded psychometric model. The mixture prior densities (overlain on the histograms) are not perfect, but they approximate the corresponding histograms much better than the Beta(2, 2) densities in Figure 6.9.*

Thanks

Random Variables

Anita Wang

2017 - 2018

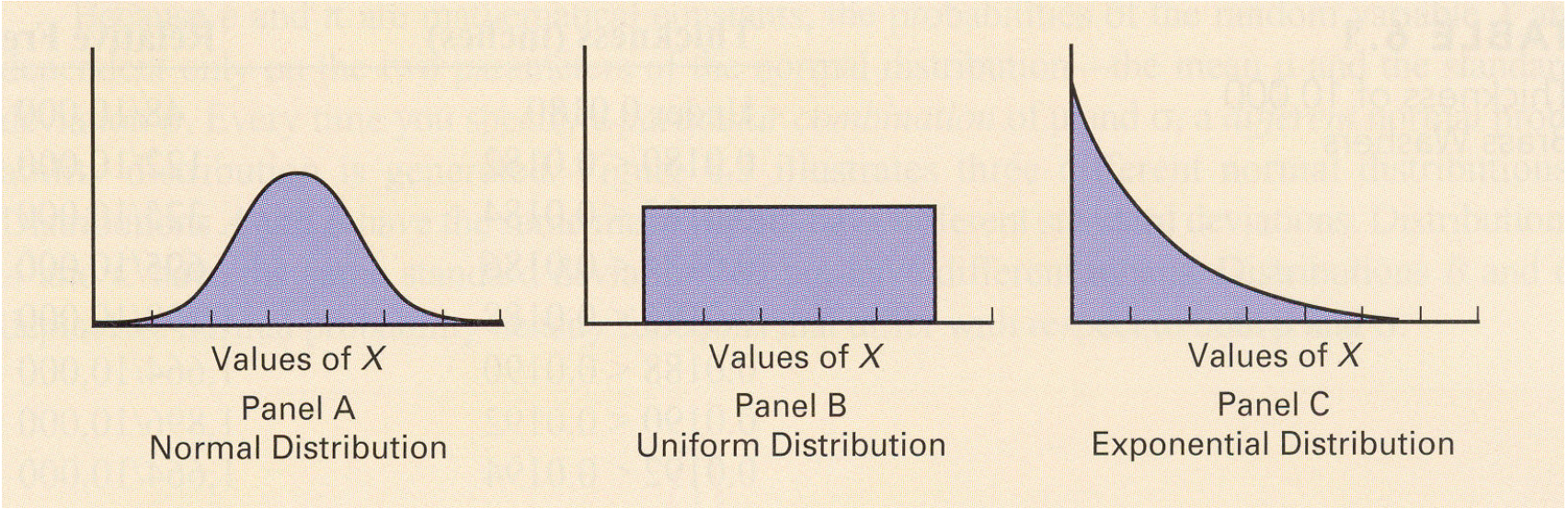
Random Variables

- Continuous random variables:
  - Uniform
  - Univariate normal
  - Gamma
  - Beta
  - Exponential

- Discrete random variables:
  - Poisson
  - Binomial

Continuous Random Variable

- A random variable is called **continuous** if the distribution function is continuous and is differentiable everywhere with the possible exception of a countable number of values.



Properties and Moments

- Properties of pdf:
  - $f(x) \geq 0$ for all x
  - $\int f(x)dx = 1$

- Expectation:

$$E(X) = \int xf(x)\,dx$$

- Variance:

$$Var(X) = \int x^2 f(x)\,dx - [E(X)]^2$$

Uniform Distribution

- The uniform distribution is used to represent a variable that is known to lie in an interval and equally likely to be found anywhere in the interval.

- $X \sim U(\alpha, \beta)$, boundaries α, β with $\beta > \alpha$
- $f(X) = \frac{1}{\beta - \alpha}$
- $E(X) = \int_{\alpha}^{\beta} \frac{1}{\beta-\alpha} x \, dx = \frac{\alpha+\beta}{2}$
- $Var(X) = \int_{\alpha}^{\beta} \frac{1}{\beta-\alpha} x^2 \, dx - \left(\frac{\alpha+\beta}{2}\right)^2 = \frac{(\beta-\alpha)^2}{12}$

Univariate Normal Distribution

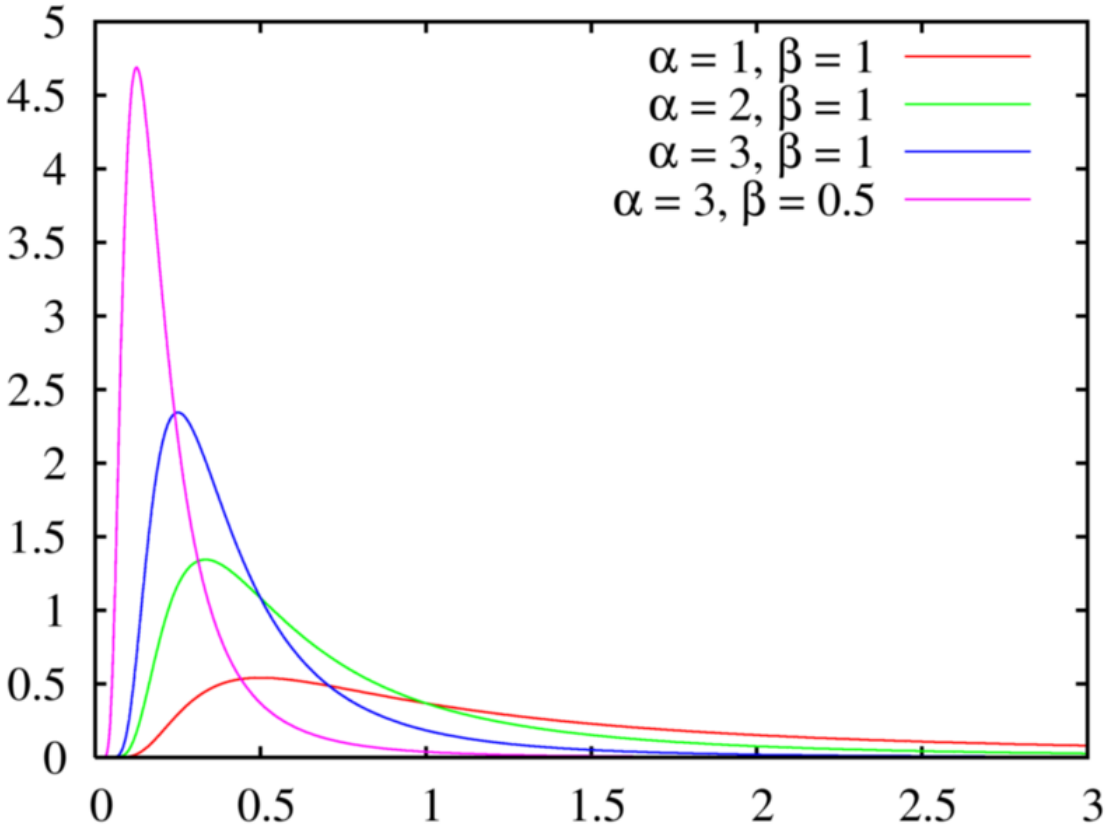
- The normal, or Gaussian, distribution is ubiquitous in statistics. Sample averages are approximately normally distributed by the central limit theorem.
- $X \sim N(\mu, \sigma^2)$, location μ, scale $\sigma > 0$
- $f(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
- $E(X) = \int_{\alpha}^{\beta} x \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)\, dx = \mu$
- $Var(X) = \int_{\alpha}^{\beta} x^2 \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)\, dx - \mu^2 = \sigma^2$

Univariate Normal Distribution

- Properties:
  - The sum of two independent normal random variables is normally distributed. If X_1 and X_2 are independent with $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
  - If $X_1|X_2 \sim N(X_2, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$. Then $X_1 \sim N(\mu_2, \sigma_1^2 + \sigma_2^2)$

Gamma Distribution

- $X \sim Gamma(\alpha, \beta)$, shape $\alpha > 0$, inverse scale $\beta > 0$



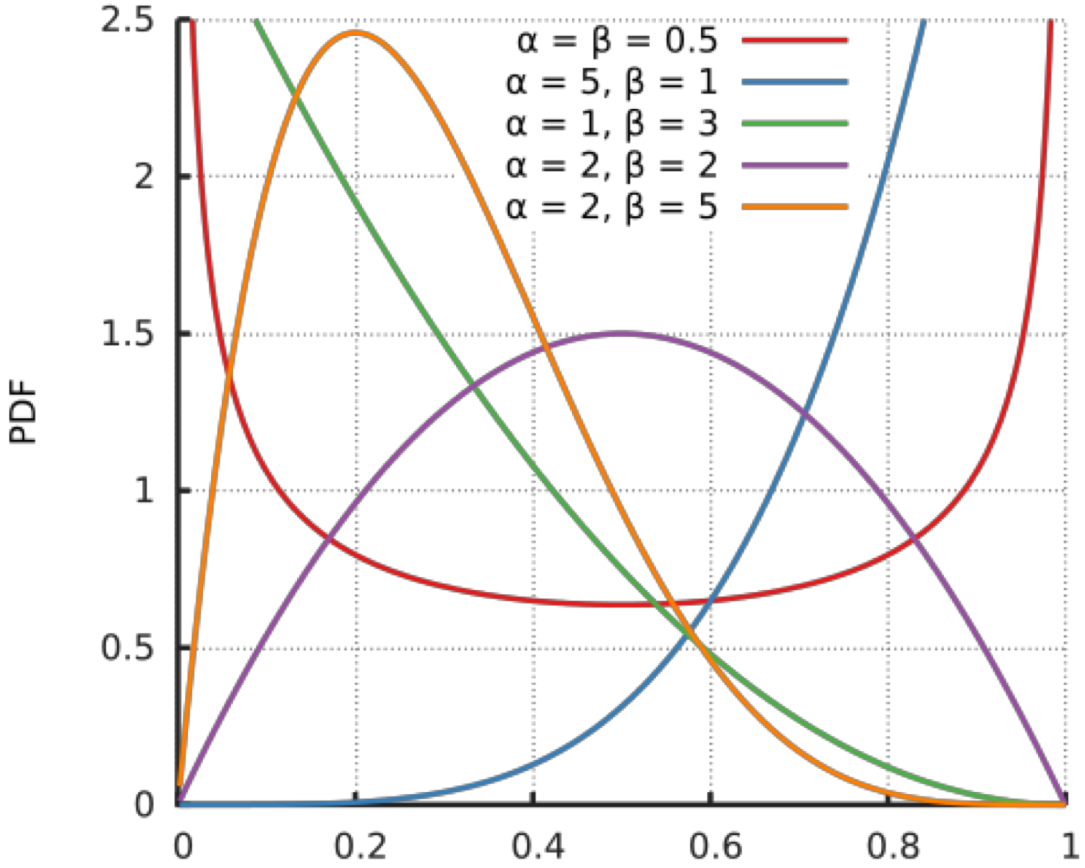
Gamma Distribution

- $f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, x > 0$
- $E(X) = \int_0^{+\infty} x \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx = \frac{\alpha}{\beta}$
- $Var(X) = \int_0^{+\infty} x^2 \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}\, dx - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}$
- Property:
  - If X_1 and X_2 are independent with Gamma(α_1, β) and Gamma(α_2, β) distributions, then

    $X_1 + X_2 \sim$ Gamma($\alpha_1 + \alpha_2$, β).

Beta Distribution

- $X \sim Beta(\alpha, \beta)$, shape $\alpha > 0$, shape $\beta > 0$

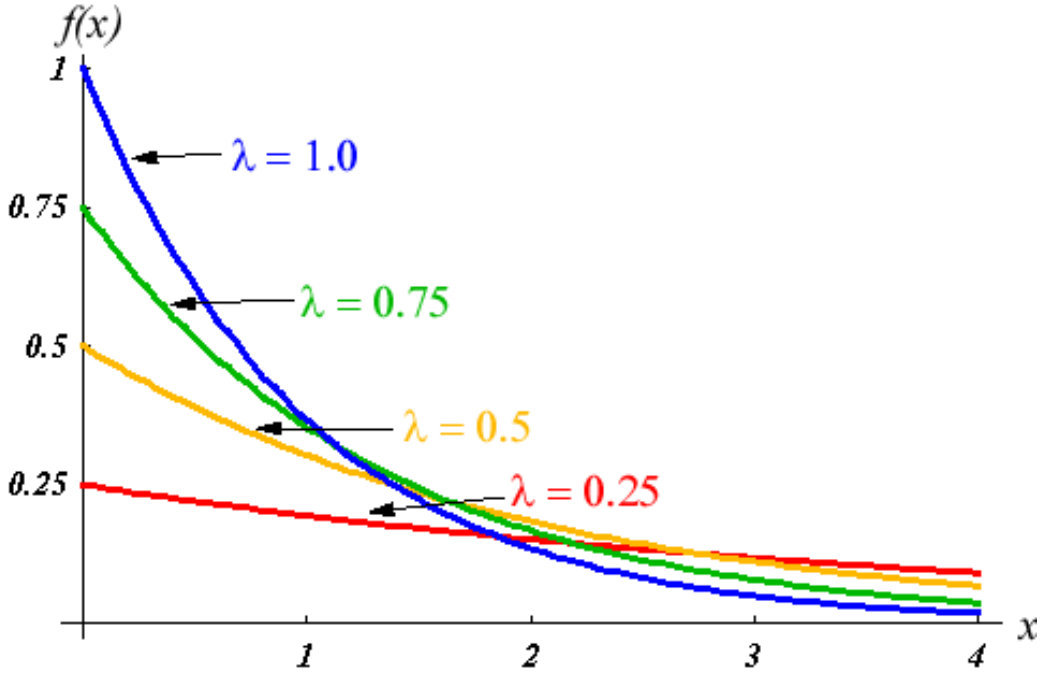


Beta Distribution

- $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, x \in [0,1], \mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$
- $E(X) = \int x \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} dx = \frac{\alpha}{\alpha+\beta}$
- $Var(X) = \int x^2 \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} dx - \left(\frac{\alpha}{\alpha+\beta}\right)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Exponential Distribution

- The exponential distribution is the distribution of waiting times for the next event in a Poisson process and is a special case of the gamma distribution with $\alpha = 1$.
- $X \sim Expon(\lambda)$, inverse scale $\lambda > 0$



Exponential Distribution

- $f(X) = \lambda e^{-\lambda x}, x > 0$
- $E(X) = \int_0^{+\infty} x\lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}$
- $Var(X) = \int_0^{+\infty} x^2\lambda e^{-\lambda x}\, dx - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$

Discrete Random Variable

- The support of a random variable is the set of numbers that are possible values of the random variable.
- A random variable is called discrete if the support contains at most a countable number of values.

Discrete Random Variable

- Properties of probability function
  - For any value x of the random variable, p(x) ≥ 0
  - The probabilities of all the events in the sample space must sum to 1, that is, $\sum_{all\ x} p(x) = 1.$
- Expectation $E(X) = \sum_{all\ x} xp(x)$
- Variance $\mathrm{Var}(X) = \sum_{all\ x} \big(x - E(X)\big)^2 p(x)$

Example

- Let Y be a discrete random variable with probability function given in the following table.

y_i	$f(y_i)$
0	0.20
1	0.15
2	0.25
3	0.35
4	0.05

- Find $E(Y)$
- Find $Var(Y)$

Example Solutions

- $E(Y) = 0\times0.20 + 1\times0.15 + 2\times0.25 + 3\times0.35 + 4\times0.05 = 1.90$
- $Var(Y) = (0-1.90)^2\times0.20 + (1-1.90)^2\times0.15 + (2-1.90)^2\times0.25 + (3-1.90)^2\times0.35 + (4-1.90)^2\times 0.05 = 1.49$

Poisson Distribution

- The Poisson distribution is commonly used to represent count data, such as the number of arrivals in a fixed time period.
- $X \sim Poisson(\lambda)$
- $p(x) = \frac{1}{x!}\lambda^x \exp(-\lambda), x = 0, 1, 2, \ldots$
- Expectation

$$E(X) = \sum_{k \geq 0} k \frac{1}{k!} \lambda^k e^{-\lambda}$$

$$E(X) = \lambda e^{-\lambda} \sum_{k \geq 1} \frac{1}{(k-1)!} \lambda^{k-1} \quad \text{as the } k = 0 \text{ term vanishes}$$

$$= \lambda e^{-\lambda} \sum_{j \geq 0} \frac{\lambda^j}{j!} \quad \text{putting } j = k - 1$$

$$= \lambda e^{-\lambda} e^{\lambda} \quad \text{Taylor Series Expansion for Exponential Function}$$

$$= \lambda$$

Poisson Distribution

- Similarly, the variance is

$$Var(X) = E(X)^2 - (EX)^2 = \lambda$$

- Property:
  - if X_1 and X_2 are independent with Poisson(λ_1) and Poisson(λ_2) distributions, then

$$X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2).$$

Binomial Distribution

- The binomial distribution is commonly used to represent the number of ‘successes’ in a sequence of n independent and identically distributed Bernoulli trials, with probability of success p in each trial.
- $X \sim Bin(n, p)$
- $p(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, \ldots n$
- Expectation

  $\mathrm{E}(\mathrm{X}) = np$
- Variance

  $Var(X) = np(1-p)$

Thanks

Review PPT

Bayesian Statistics

Anita Wang

2017 - 2018

Bayesian Theorem

prior distribution *data distribution*

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

posterior density

An Example: College Students Sleeping

- Parameter p: the proportion of American college students who sleep at least eight hours.
- A sample of 27 students is taken. In this group, 11 record that they had at least eight hours of sleep the previous night
- Discrete prior probability:

$$\begin{cases} \Pr(p = 0.2) = 0.6 \\ \Pr(p = 0.4) = 0.3 \\ \Pr(p = 0.7) = 0.1 \end{cases}$$

An Example: College Students Sleeping

The posterior probability:

$$Pr(p = 0.2|y) = \frac{Pr(p = 0.2)Pr(y|p = 0.2)}{\sum Pr(p)\Pr(y|p)}$$

$$= \frac{0.6 * \binom{27}{11}0.2^{11}0.8^{16}}{0.6\binom{27}{11}0.2^{11}0.8^{16} + 0.3\binom{27}{11}0.4^{11}0.6^{16} + 0.1\binom{27}{11}0.7^{11}0.3^{16}} = 0.089$$

$$Pr(p = 0.4|y) = \frac{Pr(p = 0.4)Pr(y|p = 0.4)}{\sum Pr(p)\Pr(y|p)}$$

$$= \frac{0.3 * \binom{27}{11}0.4^{11}0.6^{16}}{0.6\binom{27}{11}0.2^{11}0.8^{16} + 0.3\binom{27}{11}0.4^{11}0.6^{16} + 0.1\binom{27}{11}0.7^{11}0.3^{16}} = 0.909$$

$$Pr(p = 0.7|y) = \frac{Pr(p = 0.7)Pr(y|p = 0.7)}{\sum Pr(p)\Pr(y|p)}$$

$$= \frac{0.1 * \binom{27}{11}0.7^{11}0.3^{16}}{0.6\binom{27}{11}0.2^{11}0.8^{16} + 0.3\binom{27}{11}0.4^{11}0.6^{16} + 0.1\binom{27}{11}0.7^{11}0.3^{16}} = 0.002$$

Bayesian Thinking

- Parameter θ is unknown and to be estimated
- Previously, we use sample data information to estimate θ (For example, sample proportion $\hat{p}$ to estimate population proportion p)
- Bayesian thinking:
  1) Prior information of the parameter: the subject prior opinion of the distribution of the parameter
  2) Sample data information
  3) Posterior distribution: combine the information in the data with the prior distribution

Statistical Inference

Two main approaches

- Frequentist

Model parameters are fixed unknown quantities.
Randomness only in data.

  - Estimation - Maximum likelihood, method of moments
  - Confidence intervals
  - Significance testing - p-values
  - Hypothesis testing - Reject/Don't Reject H_0

Statistical Inference

- Bayesian

Model parameters are random variables. Inference is based on $P(\theta|\text{Data})$, the posterior distribution given the data.

  - Estimation - Posterior means, modes
  - Credible intervals/sets
  - Posterior probabilities

Bayes' Rule

An equivalent form omits the factor p(y), which does not depend on θ and, with fixed y, can thus be considered a constant, yielding the **unnormalized posterior density**,

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

The second term in this expression, p(y|θ), is taken here as a function of θ, not of y.

Prediction

- **Prior predictive distribution** (also called marginal distribution of y)

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

prior because it is not conditional on a previous observation of the process, and predictive because it is the distribution for a quantity that is observable.

Prediction

- Posterior predictive distribution

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

Once the data y have been observed, the unknown observable $\tilde{y}$ can be predicted. For example, $y = (y_1, y_2, \ldots, y_n)$ may be the vector of recorded weights of an object weighed n times on a scale, $\theta = (\mu, \sigma^2)$ is the prior, and $\tilde{y}$ may be the yet to be recorded weight of the object in a planned new weighing.

Simulate the Posterior Predictive Distribution

- Assuming that you can simulate from the posterior distribution of the parameter, which is usually feasible.
- To simulate the posterior predictive distribution involves two steps:

1. Simulate θ_i from $\theta|y$; $i = 1, \ldots, m$
2. Simulate $\tilde{y}_i$ from $\tilde{y}|\theta_i$ $(= \tilde{y}|\theta_i, y)$

The pairs $(\theta_i, \tilde{y}_i)$ are draws from the joint distribution $\theta, \tilde{y}|y$. Therefore the $\tilde{y}_i$ are draws from $\tilde{y}|y$.

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,

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

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

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

- This definition is formally vague since if we choose P as the class of all distributions, then P is always conjugate no matter what class of sampling distributions is used.

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.

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

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

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:

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

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

Exponential Distribution

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

Jeffreys' Priors

- Jeffreys' principle leads to defining the noninformative prior density

$$p(\theta) = [J(\theta)]^{1/2}$$

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

$$J(\theta) = E\left[\left(\frac{d \log p(y|\theta)}{d\theta}\right)^2 |\theta\right] = -E\left[\frac{d^2 \log p(y|\theta)}{d\theta^2} |\theta\right]$$

Univariate Normal with a Noninformative Prior

- Consider a vector y of n independent observations from a univariate normal distribution, $N(\mu, \sigma^2)$
- Assuming prior independence of location and scale parameters, is uniform on (μ, log σ) or,

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

- The joint posterior distribution

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

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2])$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

The marginal posterior distribution

- The marginal posterior distribution, $p(\sigma^2|y)$

$$p(\sigma^2|y) \propto \int \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu$$

$$\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \sqrt{\frac{2\pi\sigma^2}{n}}$$

$$\propto (\sigma^2)^{-\frac{n+1}{2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)$$

which is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2).$$

Joint Posterior Density

- $p(\sigma^2|y)$ is a scaled inverse-χ^2 density:

$$\sigma^2|y \sim \text{Inv} - \chi^2(n-1, s^2)$$

Therefore,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Note that this result agrees with the standard frequentist result on the sample variance.

- As we know before,

$$\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n)$$

The joint posterior density,

$$p(\mu, \sigma^2|y) \propto p(\mu|\sigma^2, y)p(\sigma^2|y)$$

Sampling from the joint posterior distribution

- Now that we have $p(\mu|\sigma^2, y)$ and $p(\sigma^2|y)$, inference on μ isn't difficult.
- One method is to use the Monte Carlo approach discussed earlier
  1. Sample σ_i^2 from $p(\sigma^2|y)$
  2. Sample μ_i from $p(\mu|\sigma_i^2, y)$

Then $\mu_1, \ldots, \mu_m$ is a sample from $p(\mu|y)$.

- Note that in this case, it is actually possible to derive the exact density of $p(\mu|y)$.

Marginal Posterior Distribution for μ

- The marginal posterior distribution

$$\mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

Therefore,

$$\frac{\mu - \bar{y}}{s/\sqrt{n}}|y \sim t_{n-1}$$

which corresponds to the standard result used for inference on a population mean

$$\frac{\bar{y} - \mu}{s/\sqrt{n}}|\mu, \sigma^2 \sim t_{n-1}$$

- The sampling distribution of the pivotal quantity $(\bar{y} - \mu)/(s/\sqrt{n})$ does not depend on the nuisance parameter σ^2, and its posterior distribution does not depend on data.

Conjugate Prior

- This has been labelled as $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2)$ distribution
- its four parameters can be identified as the location and scale of μ and the degrees of freedom and scale of σ^2
- One important thing to note is that with this prior, μ and σ^2 are dependent (i.e. $p(\mu|\sigma^2)$ is a function of σ^2, for example, if σ^2 is large, then a high-variance prior distribution is induced on μ
- This has a different feel from the standard frequentist analysis where $\bar{y}$ and s^2 are independent.

The Posterior Density

- The posterior density satisfies

$$p(\mu,\sigma^2|y) \propto \frac{1}{\sigma}\frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2]\right)$$

$$\times\frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\bar{y}-\mu)^2])$$

$$\propto \frac{1}{\sigma}\frac{1}{(\sigma^2)^{\frac{\nu_n}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2+\kappa_n(\mu-\mu_n)^2]\right)$$

The posterior distribution is $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_n,\frac{\sigma_n^2}{\kappa_n};\nu_n,\sigma_n^2)$

The Posterior Density

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2$$

The parameters of the posterior distribution combine the prior information and the information contained in the data. For example μ_n is a weighted average of the prior mean and the sample mean, with weights determined by the relative precision of the two pieces of information.

The Conditional Posterior Distribution $p(\mu|\sigma^2, y)$

- By using that $p(\mu|\sigma^2, y) \propto p(\mu, \sigma^2|y)$ with σ as a constant, we get

$$\mu|\sigma^2, y \sim N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

Note that the mean and variance can be written as

$$\mu_n = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} \qquad \sigma_n^2 = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}$$

which matches with the fixed variance case discuss earlier.

The Marginal Posterior Distribution $p(\sigma^2|y)$

- $p(\sigma^2|y)$

$$\sigma^2|y \sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)$$

This can be seen by the same way $p(\sigma^2|y)$ was shown in the non-informative prior case or by recognizing the $\text{N} - \text{Inv} - \chi^2$ form of the joint density.

- $p(\mu|y)$

As mentioned before, this can be determined by simulation (see in the next slide). In this case an exact answer can be determined by integrating out σ^2 from the joint density (as in the non-informative case), we get

$$\mu|y \sim t_{\nu_n}(\mu_n, \frac{\sigma_n^2}{\kappa_n})$$

Simulation of $p(\mu|y)$

- we first draw σ^2 from its marginal posterior distribution $p(\sigma^2|y)$,

$$\sigma^2|y \sim \text{Inv} - \chi^2(\nu_n, \sigma_n^2)$$

- then draw μ from its normal conditional posterior distribution $p(\mu|\sigma^2, y)$

$$\mu|\sigma^2, y \sim N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

using the simulated value of σ^2.

The Prior and Posterior Distribution

- The conjugate prior distribution

Dirichlet: a multivariate generalization of the beta distribution

$$p(\theta|\alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}$$

where $\theta_j \in (0,1)$ and $\sum \theta_j = 1$

- The posterior distribution

The resulting posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$.

Thanks