Outline of the chapter 2

- 2.1 Binomial model (repeated experiment with binary outcome)
- 2.2 Posterior as compromise between data and prior information
- 2.3 Posterior summaries
- 2.4 Informative prior distributions (skip exponential families and sufficient statistics)
- 2.5 Gaussian model with known variance
- 2.6 Other single parameter models
  - the normal distribution with known mean but unknwon variance is the most important
  - glance through Poisson and exponential
- 2.7 glance through this example, which illustrates benefits of prior information, no need to read all the details (it's quite long example)
- 2.8 Noninformative and weakly informative priors

Binomial: known θ

- Probability of event 1 in trial is θ

Binomial: known θ

- Probability of event 1 in trial is θ
- Probability of event 2 in trial is $1 - \theta$

Binomial: known θ

- Probability of event 1 in trial is θ
- Probability of event 2 in trial is $1 - \theta$
- Probability of seveal events in independent trials is e.g. $\theta\theta(1-\theta)\theta(1-\theta)(1-\theta)\dots$

Binomial: known θ

- Probability of event 1 in trial is θ
- Probability of event 2 in trial is $1 - \theta$
- Probability of seveal events in independent trials is e.g. $\theta\theta(1-\theta)\theta(1-\theta)(1-\theta)\dots$
- If there are n trials and we don't care about the order of the events, then the probability that event 1 happens y times is

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

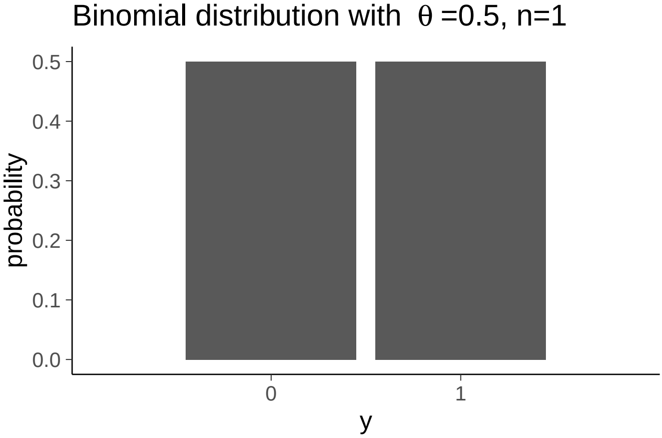
Binomial: known θ

- Observation model (function of y, discrete)

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

- Observation model (function of y, discrete)

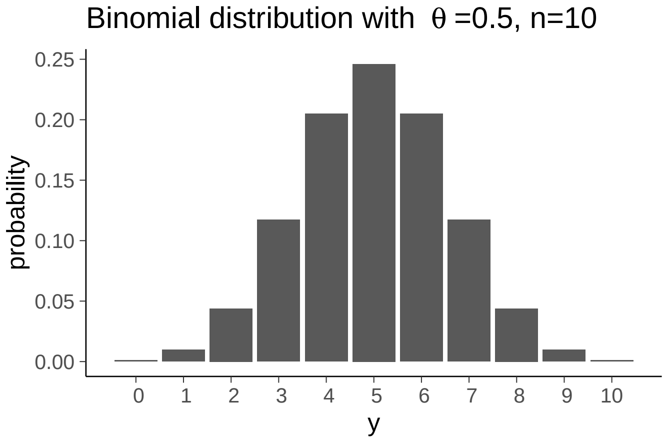
$$p(y|\theta, n) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$



Binomial: known θ

- Observation model (function of y, discrete)

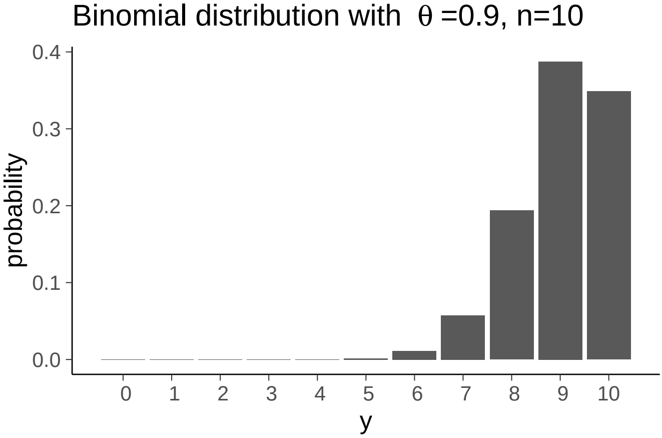
$$p(y|\theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Binomial: known θ

- Observation model (function of y, discrete)

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



Binomial: unknown θ

- Posterior with Bayes rule (function of θ, continuous)

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

Binomial: unknown θ

- Posterior with Bayes rule (function of θ, continuous)

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

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

Binomial: unknown θ

- Posterior with Bayes rule (function of θ, continuous)

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

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

- Start with uniform prior

$$p(\theta|n, M) = p(\theta|M) = 1, \text{ kun } 0 \leq \theta \leq 1$$

Binomial: unknown θ

- Posterior with Bayes rule (function of θ, continuous)

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

  where $p(y|n,M) = \int p(y|\theta,n,M)p(\theta|n,M)d\theta$
- Start with uniform prior

$$p(\theta|n,M) = p(\theta|M) = 1, \text{ kun } 0 \leq \theta \leq 1$$

- Then

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

$$= \frac{1}{Z}\theta^y(1-\theta)^{n-y}$$

Binomial: unknwon θ

- Normalization term Z (constant given y)

$$Z = p(y|n, M) = \int_0^1 \theta^y (1-\theta)^{n-y} d\theta = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

- Normalisation term has Beta function form
  - when integarted over $(0, 1)$ the result can presented with Gamma functions
  - with integers $\Gamma(n) = (n-1)!$
  - for large integers even this is challenging and usually $\log \Gamma(\cdot)$ is computed instead of $\Gamma(\cdot)$

Binomial: unknwon θ

- Posterior is

$$p(\theta|y, n, M) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y}(1-\theta)^{n-y},$$

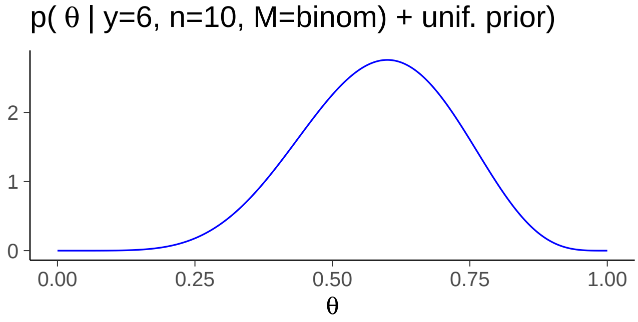
Binomial: unknwon θ

- Posterior is

$$p(\theta|y,n,M) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y},$$

which is called Beta distribution

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



Binomial: computation

- R
  - density `dbeta`
  - CDF `pbeta`
  - quantile `qbeta`
  - random number `rbeta`
- Python
  - `from scipy.stats import beta`
  - density `beta.pdf`
  - CDF `beta.cdf`
  - prctile `beta.ppf`
  - random number `beta.rvs`

Binomial: computation*

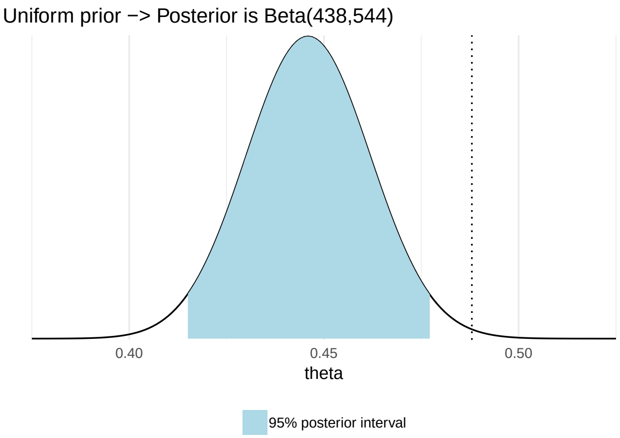
- Beta CDF not trivial to compute
- For example, `pbeta` in R uses a continued fraction with weighting factors and asymptotic expansion
- Laplace developed normal approximation (Laplace approximation), because he didn't know how to compute Beta CDF

Placenta previa

- Probability of a girl birth given placenta previa (BDA3 p. 37)
  - 437 girls and 543 boys have been observed
  - is the ratio 0.445 different from the population average 0.485?

Placenta previa

- Probability of a girl birth given placenta previa (BDA3 p. 37)
  - 437 girls and 543 boys have been observed
  - is the ratio 0.445 different from the population average 0.485?



Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1 | \theta, y, n, M)$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1|y, n, M) = \int_0^1 p(\tilde{y} = 1|\theta, y, n, M)p(\theta|y, n, M)d\theta$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$
\begin{aligned}
p(\tilde{y} = 1 | y, n, M) &= \int_0^1 p(\tilde{y} = 1 | \theta, y, n, M) p(\theta | y, n, M) d\theta \\
&= \int_0^1 \theta p(\theta | y, n, M) d\theta
\end{aligned}
$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$
\begin{aligned}
p(\tilde{y}=1|y,n,M) &= \int_0^1 p(\tilde{y}=1|\theta,y,n,M)p(\theta|y,n,M)d\theta \\
&= \int_0^1 \theta p(\theta|y,n,M)d\theta \\
&= \mathrm{E}[\theta|y]
\end{aligned}
$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$
\begin{aligned}
p(\tilde{y} = 1|y, n, M) &= \int_0^1 p(\tilde{y} = 1|\theta, y, n, M)p(\theta|y, n, M)d\theta \\
&= \int_0^1 \theta p(\theta|y, n, M)d\theta \\
&= \mathrm{E}[\theta|y]
\end{aligned}
$$

- With uniform prior

$$
\mathrm{E}[\theta|y] = \frac{y+1}{n+2}
$$

Predictive distribution – Effect of integration

- Predictive distribution for new $\tilde{y}$ (discrete)

$$\begin{aligned}
p(\tilde{y} = 1 | y, n, M) &= \int_0^1 p(\tilde{y} = 1 | \theta, y, n, M) p(\theta | y, n, M) d\theta \\
&= \int_0^1 \theta p(\theta | y, n, M) d\theta \\
&= \mathrm{E}[\theta | y]
\end{aligned}$$

- With uniform prior

$$\mathrm{E}[\theta | y] = \frac{y+1}{n+2}$$

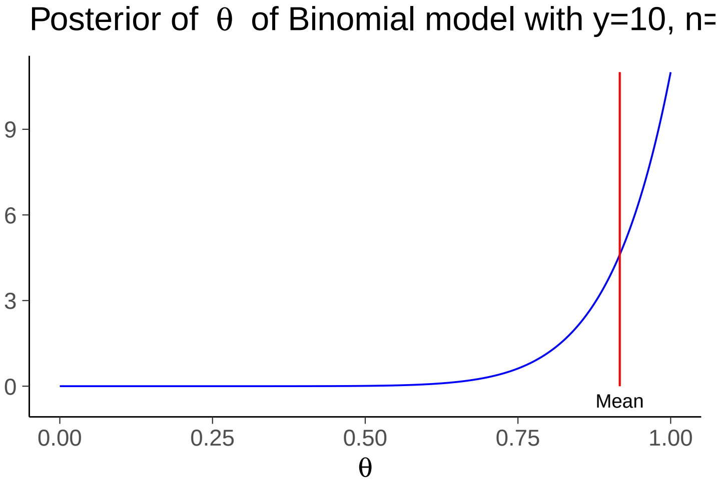
- Extreme cases

$$\begin{aligned}
p(\tilde{y} = 1 | y = 0, n, M) &= \frac{1}{n+2} \\
p(\tilde{y} = 1 | y = n, n, M) &= \frac{n+1}{n+2}
\end{aligned}$$

- cf. maximum likelihood

Benefits of integration

Example: $n = 10, y = 10$



Predictive distribution

- Prior predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1|M) = \int_0^1 p(\tilde{y} = 1|\theta, y, n, M)p(\theta|M)d\theta$$

- Posterior predictive distribution for new $\tilde{y}$ (discrete)

$$p(\tilde{y} = 1|y, n, M) = \int_0^1 p(\tilde{y} = 1|\theta, y, n, M)p(\theta|y, n, M)d\theta$$

Justification for uniform prior

- $p(\theta|M) = 1$ if
  1) we want the prior predictive distribution to be uniform

$$p(y|n, M) = \frac{1}{n+1}, \quad y = 0, \dots, n$$

    - nice justification as it is based on observables y and n

Justification for uniform prior

- $p(\theta|M) = 1$ if
  1) we want the prior predictive distribution to be uniform

$$p(y|n, M) = \frac{1}{n+1}, \quad y = 0, \ldots, n$$

  - nice justification as it is based on observables y and n

  2) we think all values of θ are equally likely

Priors

- Conjugate prior (BDA3 p. 35)
- Noninformative prior (BDA3 p. 51)
- Proper and improper prior (BDA3 p. 52)
- Weakly informative prior (BDA3 p. 55)
- Informative prior (BDA3 p. 55)
- Prior sensitivity (BDA3 p. 38)

Conjugate prior

- Prior and posterior have the same form
  - only for exponential family distributions (plus for some irregular cases)
- Used to be important for computational reasons, and still sometimes used for special models to allow partial analytic marginalization (Ch 3)
  - with Hamiltonian Monte carlo used e.g. in Stan no any computational benefit

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

$$p(\theta|y, n, M) \propto \theta^{y}(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Beta prior for Binomial model

- Prior

$$\text{Beta}(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

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

Beta prior for Binomial model

- Prior

$$\mathrm{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior

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

Beta prior for Binomial model

- Prior

$$\mathrm{Beta}(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

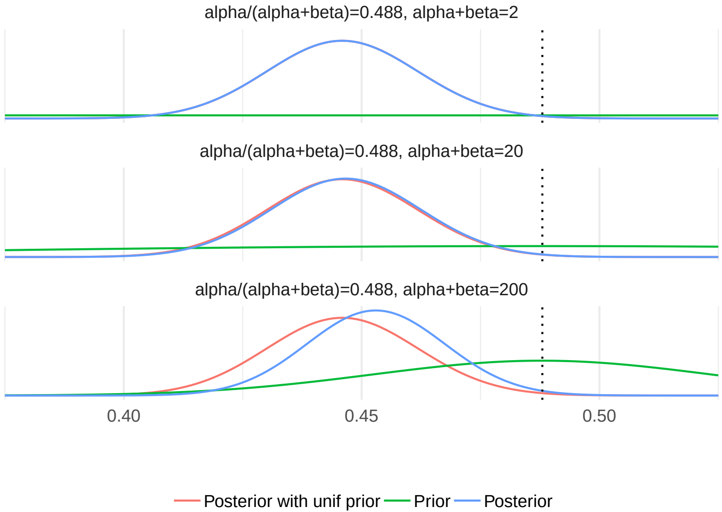
- Posterior

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

- $(\alpha-1)$ and $(\beta-1)$ can considered to be number of prior observations
- Uniform prior when $\alpha = 1$ ja $\beta = 1$

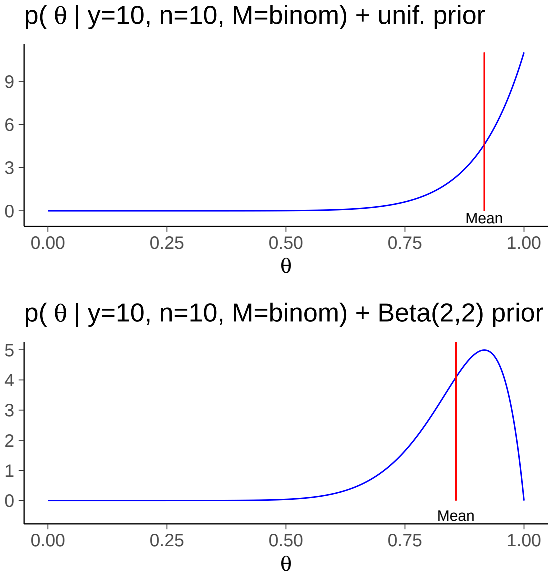
Placenta previa

- Beta prior centered on population average 0.485



Benefits of integration and prior

Example: $n = 10, y = 10$ - uniform vs Beta(2,2) prior



Beta prior for Binomial model

- Posterior

$$p(\theta|y, n, M) = \text{Beta}(\theta|\alpha + y, \beta + n - y)$$

- Posterior mean

$$\text{E}[\theta|y] = \frac{\alpha + y}{\alpha + \beta + n}$$

  - combination prior and likelihood information
  - kun $n \to \infty$, $\text{E}[\theta|y] \to y/n$

Beta prior for Binomial model

- Posterior

$$p(\theta|y,n,M) = \mathrm{Beta}(\theta|\alpha+y,\beta+n-y)$$

- Posterior mean

$$\mathrm{E}[\theta|y] = \frac{\alpha+y}{\alpha+\beta+n}$$

  - combination prior and likelihood information
  - kun $n \to \infty$, $\mathrm{E}[\theta|y] \to y/n$

- Posterior variance

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

  - decreases when n increases
  - when $n \to \infty$, $\mathrm{Var}[\theta|y] \to 0$

Noninformative prior, proper and imporper prior

- Vague, flat, diffuse of noninformative
  - try to “to let the data speak for themselves”
  - flat is not non-informative
  - flat can be stupid
  - making prior flat somewhere can make it non-flat somewhere else
- Proper prior has $\int p(\theta) = 1$
- Improper prior density doesn’t have a finite integral
  - the posterior can still sometimes be proper

Weakly informative priors

- Weakly informative priors produce computationally better behaving posteriors
    - quite often there's at least some knowledge about the scale
    - useful also if there's more information from previous observations, but not certain how well that information is applicable in a new case uncertainty

Weakly informative priors

- Weakly informative priors produce computationally better behaving posteriors
  - quite often there's at least some knowledge about the scale
  - useful also if there's more information from previous observations, but not certain how well that information is applicable in a new case uncertainty
- Construction
  - Start with some version of a noninformative prior distribution and then add enough information so that inferences are constrained to be reasonable.
  - Start with a strong, highly informative prior and broaden it to account for uncertainty in one's prior beliefs and in the applicability of any historically based prior distribution to new data.
- Stan team prior choice recommendations https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Example of informative prior

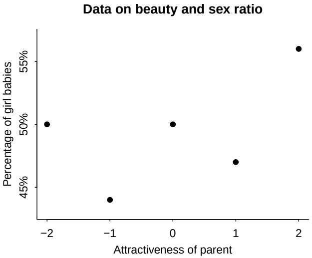
- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate

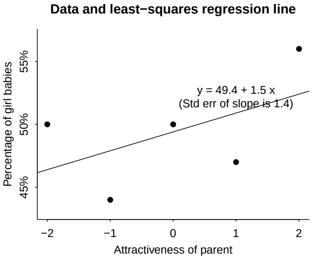
Example of informative prior

- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate
- There was a study on the percentage of girl births among parents in attractiveness categories 1–5 (assessed by interviewers in a face-to-face survey)

Example of informative prior

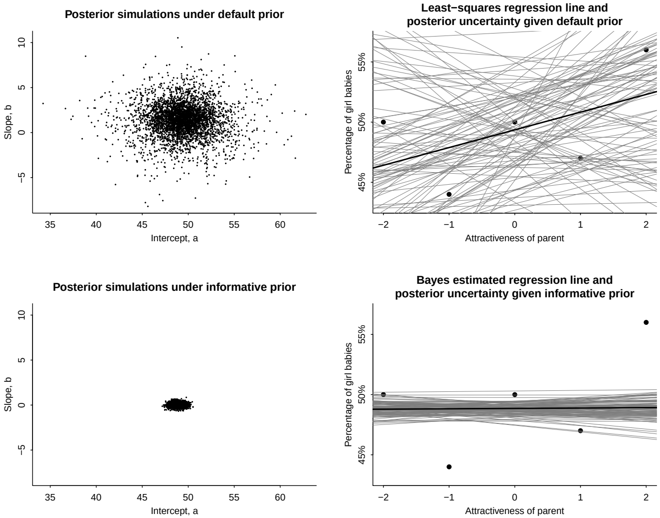
- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate
- There was a study on the percentage of girl births among parents in attractiveness categories 1–5 (assessed by interviewers in a face-to-face survey)





Example of informative prior

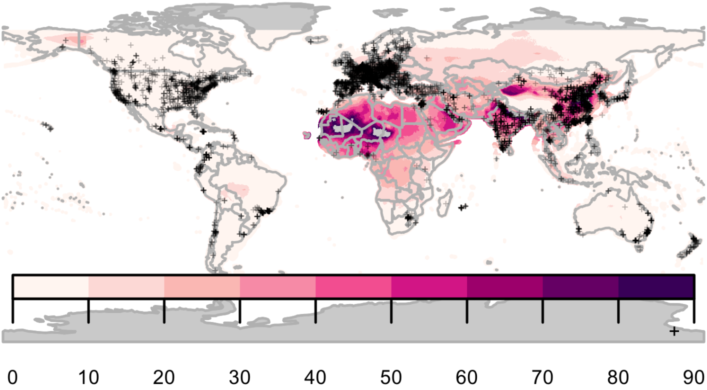
- The percentage of girl births is remarkably stable at about 48.8% girls, rarely varying by more than 0.5% from this rate



Example of weakly informative prior

- Gabry et al (2017). Visualization in Bayesian workflow.
  - Estimation of human exposure to air pollution from particulate matter measuring less than 2.5 microns in diameter ($PM_{2.5}$)
  - A recent report estimated that $PM_{2.5}$ is responsible for three million deaths worldwide each year (Shaddick et al, 2017)

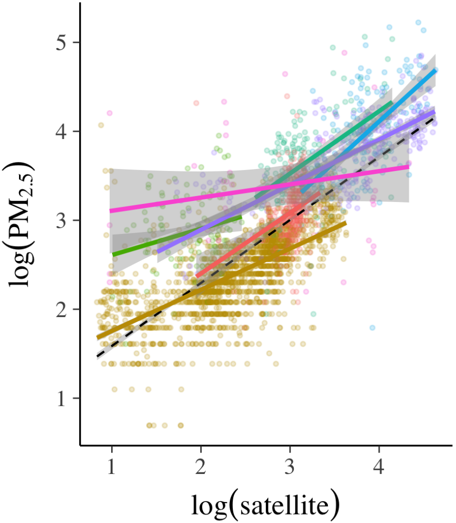
- Gabry et al (2017). Visualization in Bayesian workflow.



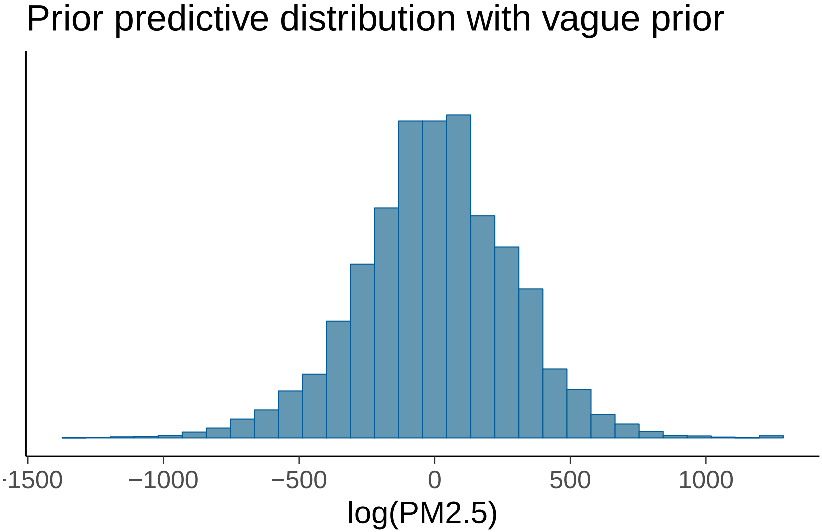
Satellite estimates and ground monitor locations

Example of weakly informative prior

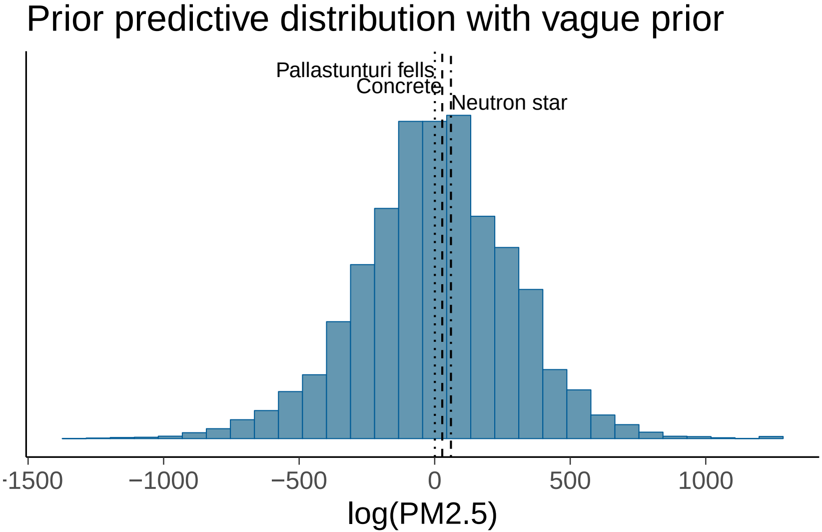
- Gabry et al (2017). Visualization in Bayesian workflow.



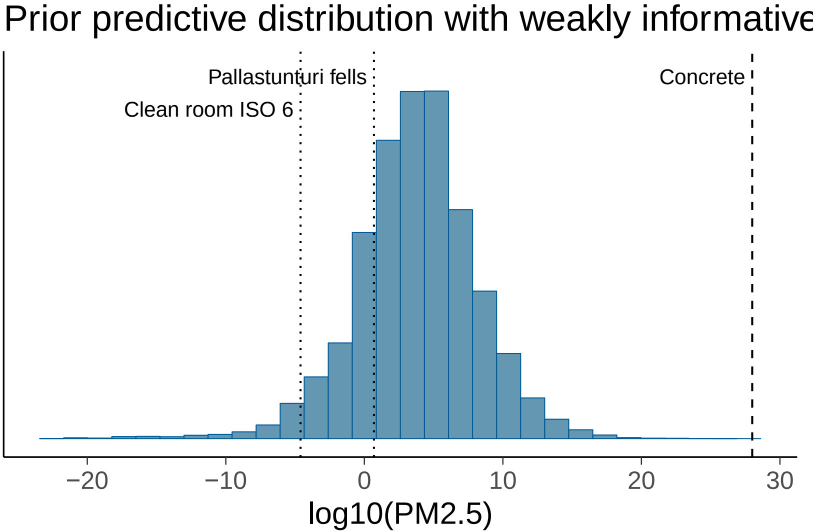
Example of weakly informative prior



Example of weakly informative prior



Example of weakly informative prior



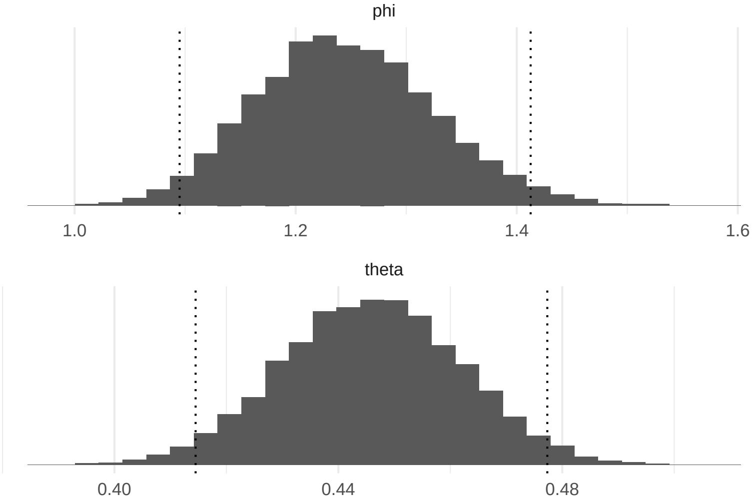
Effect of incorrect priors?

- Introduce bias, but often still produce smaller estimation error because the variance is reduced
  - bias-variance tradeoff

Sufficient statistics*

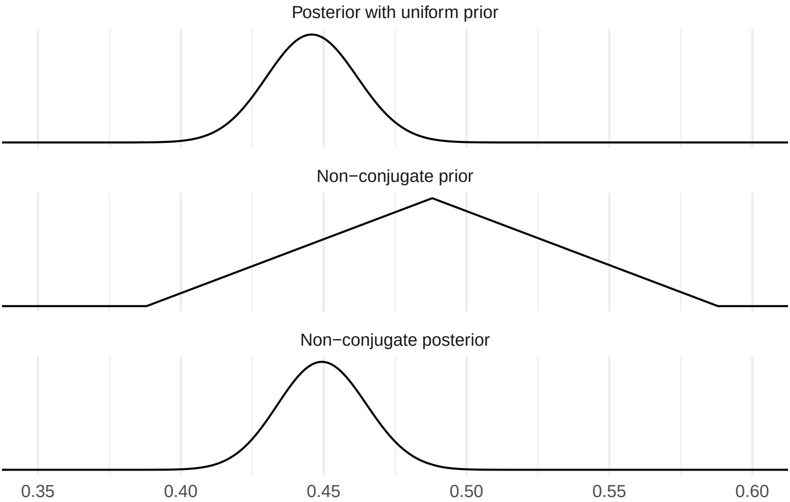
- The quantity $t(y)$ is said to be a *sufficient statistic* for θ, because the likelihood for θ depends on the data y only through the value of $t(y)$.

- demo2_3: Simulate samples from Beta(438,544), and draw a histogram of θ and OR with quantilesable.



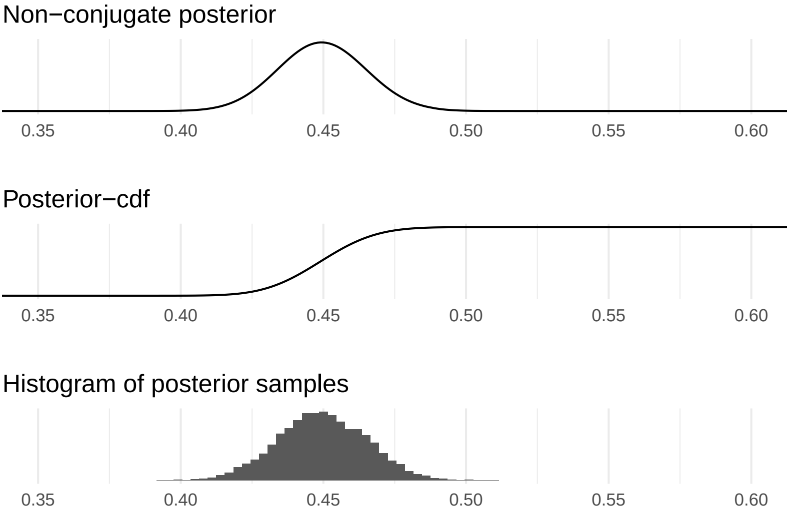
Posterior visualisation and inference demos

- demo2_4: Compute posterior distribution in a grid.



Posterior visualisation and inference demos

- demo2_4: Sample using the inverse-cdf method.



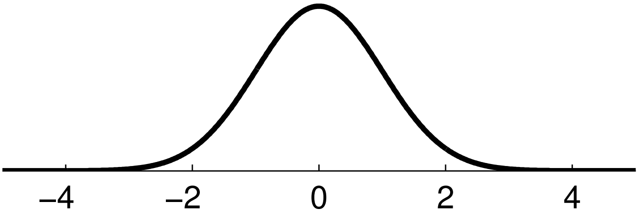
Algae status is monitored in 274 sites at Finnish lakes and rivers. The observations for the 2008 algae status at each site are presented in file algae.mat ('0': no algae, '1': algae present). Let π be the probability of a monitoring site having detectable blue-green algae levels.

- Use a binomial model for observations and a beta(2,10) prior.
- What can you say about the value of the unknown π?
- Experiment how the result changes if you change the prior.

Normal / Gaussian

- Observations y real valued
- Mean θ and variance σ^2 (or deviation σ)
  (first assume σ^2 known)

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$
$$y \sim \mathrm{N}(\theta, \sigma^2)$$



Reasons to use Normal distribution

- Normal distribution often justified based on central limit theorem
- More often used due to the computational convenience or tradition

Central limit theorem*

- De Moivre, Laplace, Gauss, Chebysev, Liapounov, Markov, et al.
- Given certain conditions sum (and mean) of random variables approach Gaussian distribution as d $n \to \infty$
- Problems
  - does not hold for all distributions, e.g., Cauchy
  - may require large n,
    e.g. Binomial, when θ close to 0 or 1
  - does not hold if one the variables has much larger scale

Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

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

Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

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

$$\exp(a)\exp(b) = \exp(a+b)$$

Normal distribution - conjugate prior for θ

- Assume σ^2 known

Likelihood $$p(y|\theta) \propto \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right)$$

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

$$\exp(a)\exp(b) = \exp(a+b)$$

Posterior $$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

Normal distribution - conjugate prior for θ

- Posterior (see ex 2.14a)

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$
$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

$$\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2), \quad \text{where} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{ja} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

Normal distribution - conjugate prior for θ

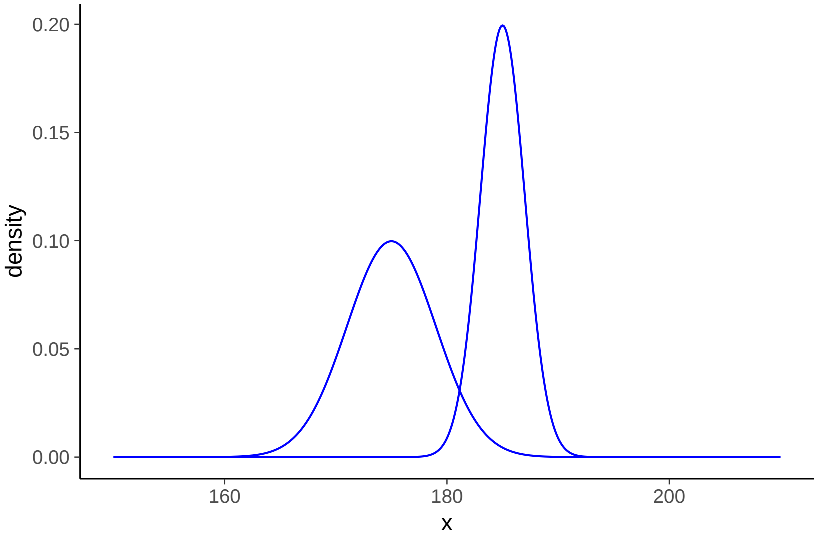
- Posterior (see ex 2.14a)

$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$
$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

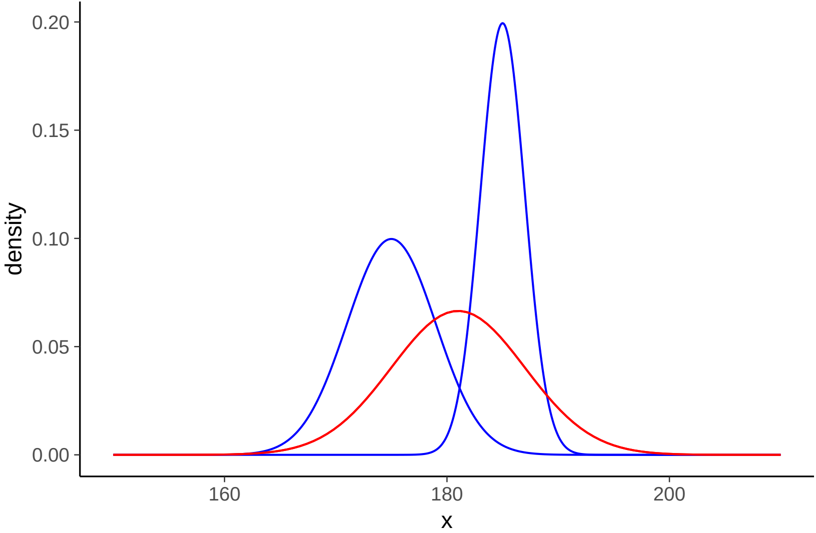
$$\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2), \quad \text{where} \quad \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \quad \text{ja} \quad \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

- 1/variance = precision
- Posterior precision = prior precision + data precision
- Posterior mean is precision weighted mean

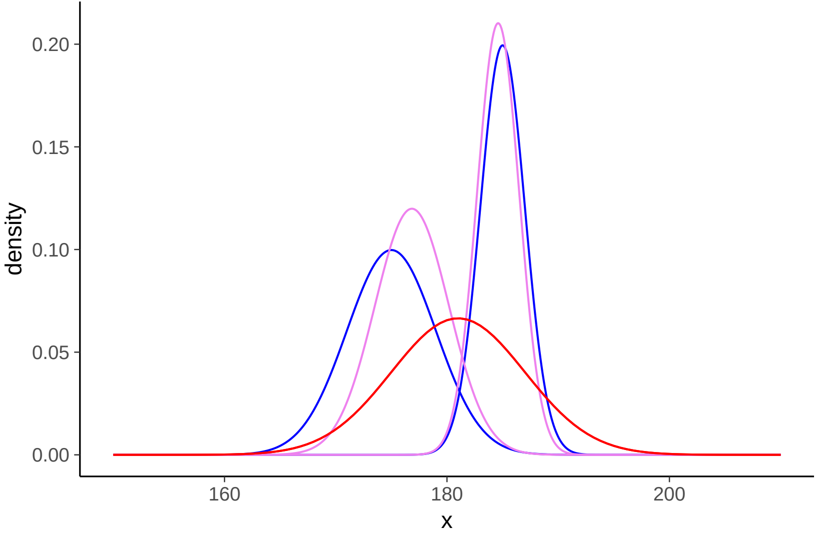
Normal distribution - example



Normal distribution - example



Normal distribution - example



Normal distribution - conjugate prior for θ

- Several observations – use chain rule

Normal distribution - conjugate prior for θ

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta|y) = \mathrm{N}(\theta|\mu_n, \tau_n^2)$$

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

- If $\tau_0^2 = \sigma^2$, prior corresponds to one virtual observation with value μ_0

Normal distribution - conjugate prior for θ

- Several observations $y = (y_1, \ldots, y_n)$

$$p(\theta|y) = \mathrm{N}(\theta|\mu_n, \tau_n^2)$$

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

- If $\tau_0^2 = \sigma^2$, prior corresponds to one virtual observation with value μ_0
- If $\tau_0 \to \infty$ when n fixed
  or if $n \to \infty$ when τ_0 fixed

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

Normal distribution - conjugate prior for θ

- Posterior predictive distribution

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

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

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

- Predictive variance = observation model variance σ^2 + posterior variance τ_1^2