UPPSALA UNIVERSITET

- Posterior distributions
- Predictive distributions
- Prior distributions
- Demo
- The Normal model

Bayesian Statistics and Data Analysis
Lecture 2

Måns Magnusson
Department of Statistics, Uppsala University
Thanks to Aki Vehtari, Aalto University

UPPSALA UNIVERSITET

Binomial: known θ

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

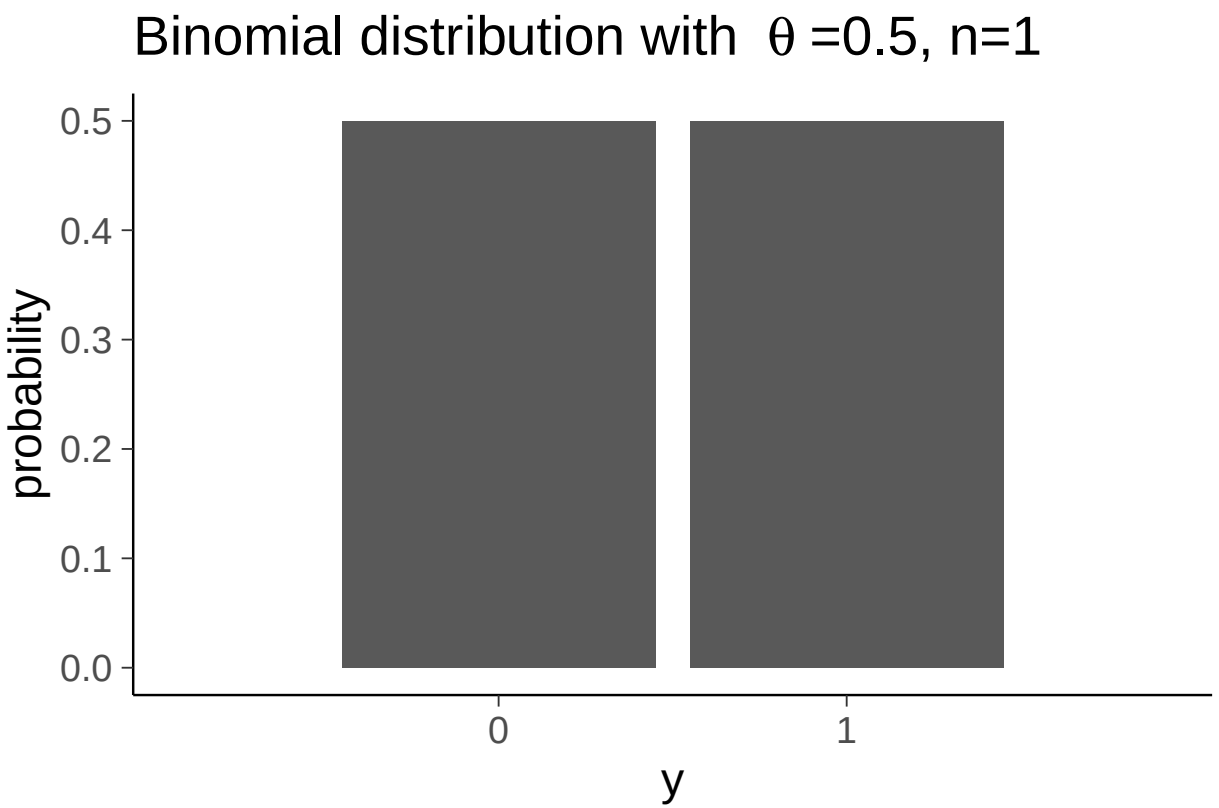
$$p(y|\theta, n) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

UPPSALA UNIVERSITET

Binomial: known θ

- Observation model (function of y, discrete)

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

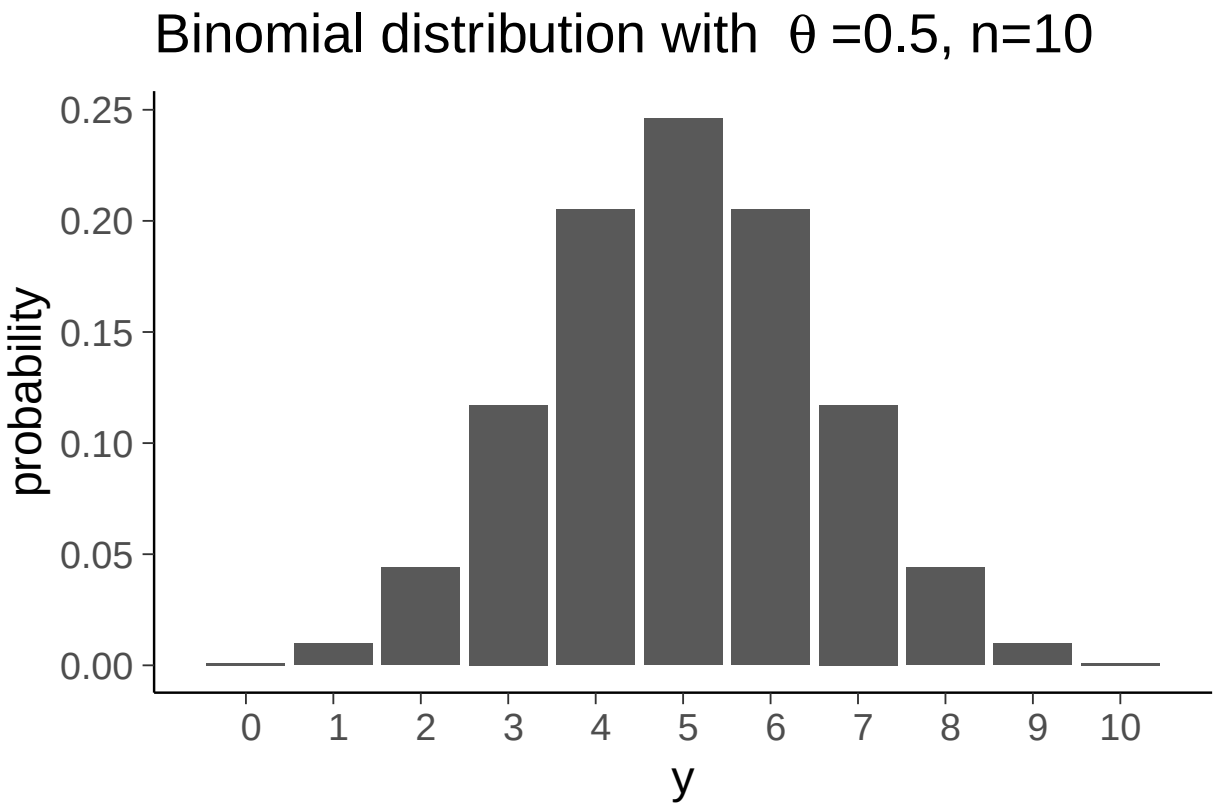


UPPSALA UNIVERSITET

Binomial: known θ

- Observation model (function of y, discrete)

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



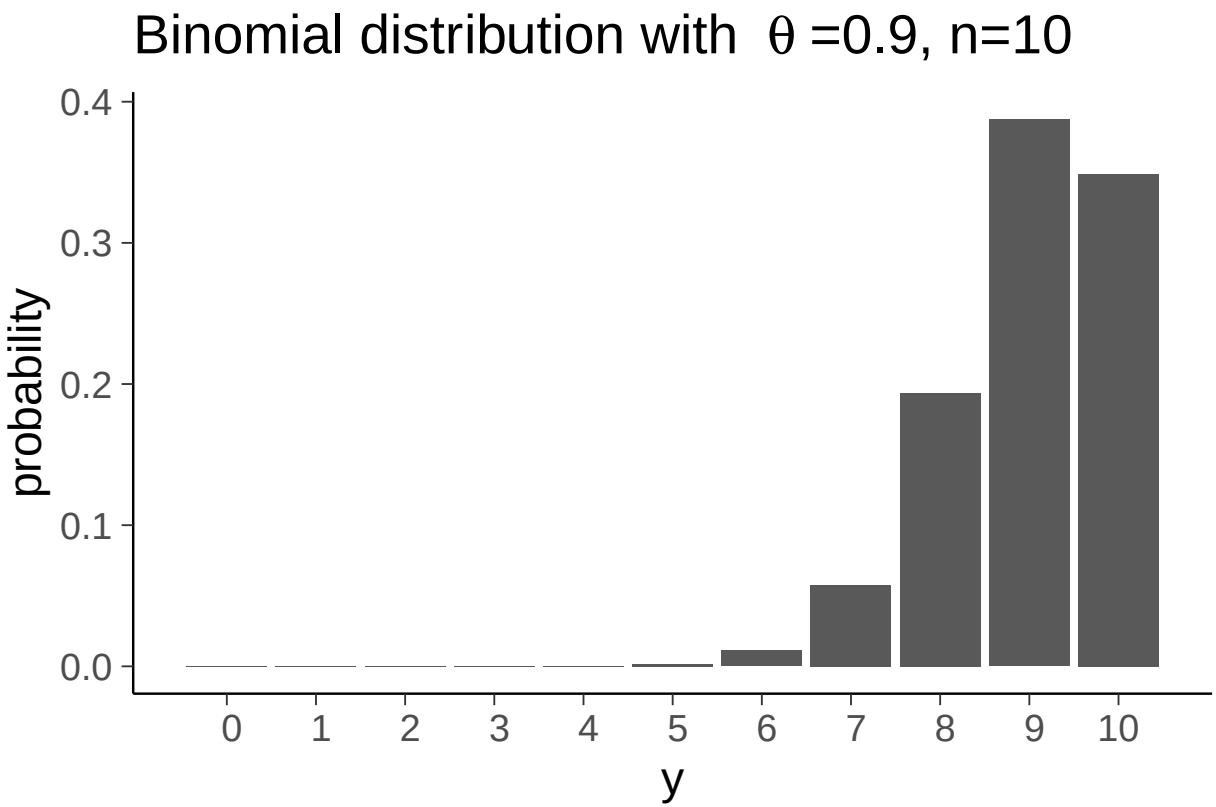
$p(y|n = 10, \theta = 0.5)$: 0.00 0.01 0.04 0.12 0.21 0.25 0.21 0.12 0.04 0.01 0.00

UPPSALA UNIVERSITET

Binomial: known θ

- Observation model (function of y, discrete)

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



$p(y|n = 10, \theta = 0.9)$: 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.06 0.19 0.39 0.35

UPPSALA UNIVERSITET

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

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

- Then

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

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

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

UPPSALA UNIVERSITET

Binomial: unknown θ

- Normalization term Z (constant given y)

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

- Normalisation term has *Beta* function form
  - when integrated 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 $\Gamma(\cdot)$

UPPSALA UNIVERSITET

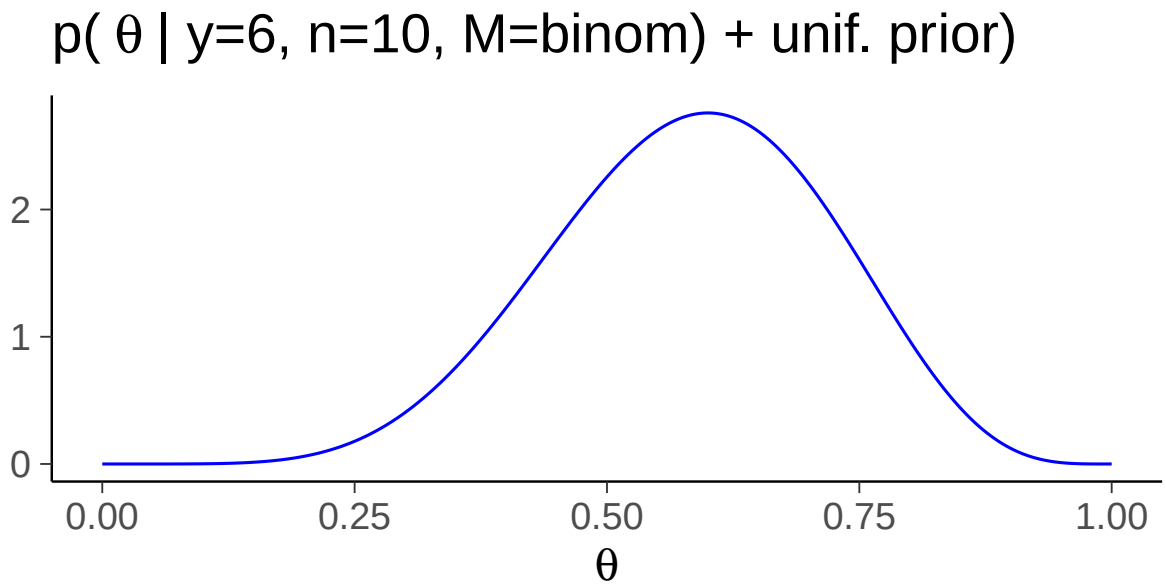
Binomial: unknown θ

- 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

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



UPPSALA UNIVERSITET

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`

UPPSALA UNIVERSITET

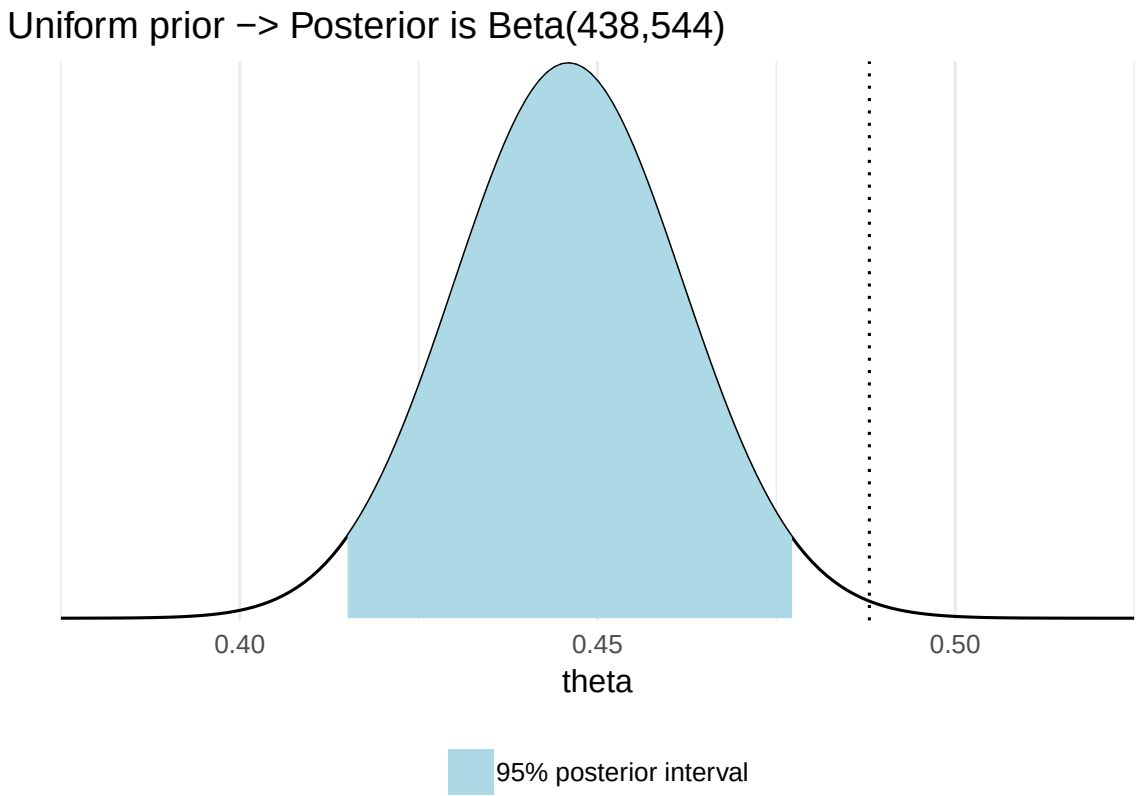
Binomial: computation*

- 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

UPPSALA UNIVERSITET

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?



UPPSALA UNIVERSITET

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

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

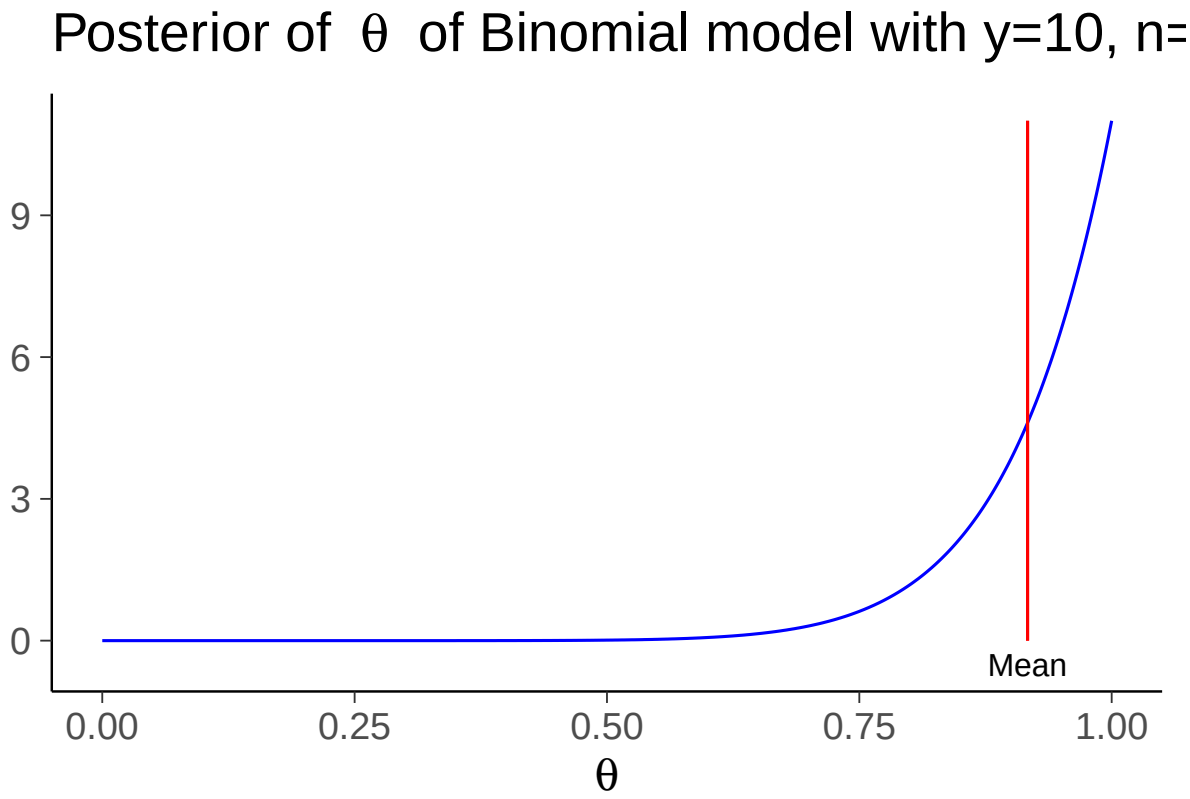
$$p(\tilde{y} = 1|y = n, n, M) = \frac{n+1}{n+2}$$

- cf. maximum likelihood

UPPSALA UNIVERSITET

Benefits of integration

Example: $n = 10, y = 10$



UPPSALA UNIVERSITET

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)

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

UPPSALA UNIVERSITET

Justification for uniform prior

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

$$p(\tilde{y} = 1|n = 0, M) = \frac{1}{2}$$

    - nice justification as it is based on observables y and n
  2) we think all values of θ are equally likely

UPPSALA UNIVERSITET

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)

UPPSALA UNIVERSITET

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
- Still used for special models to allow partial analytic marginalization (Ch 3)
  - with dynamic Hamiltonian Monte Carlo used e.g. in Stan no any computational benefit

UPPSALA UNIVERSITET

Beta prior for Binomial model

- Prior

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

- Posterior

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

after normalization

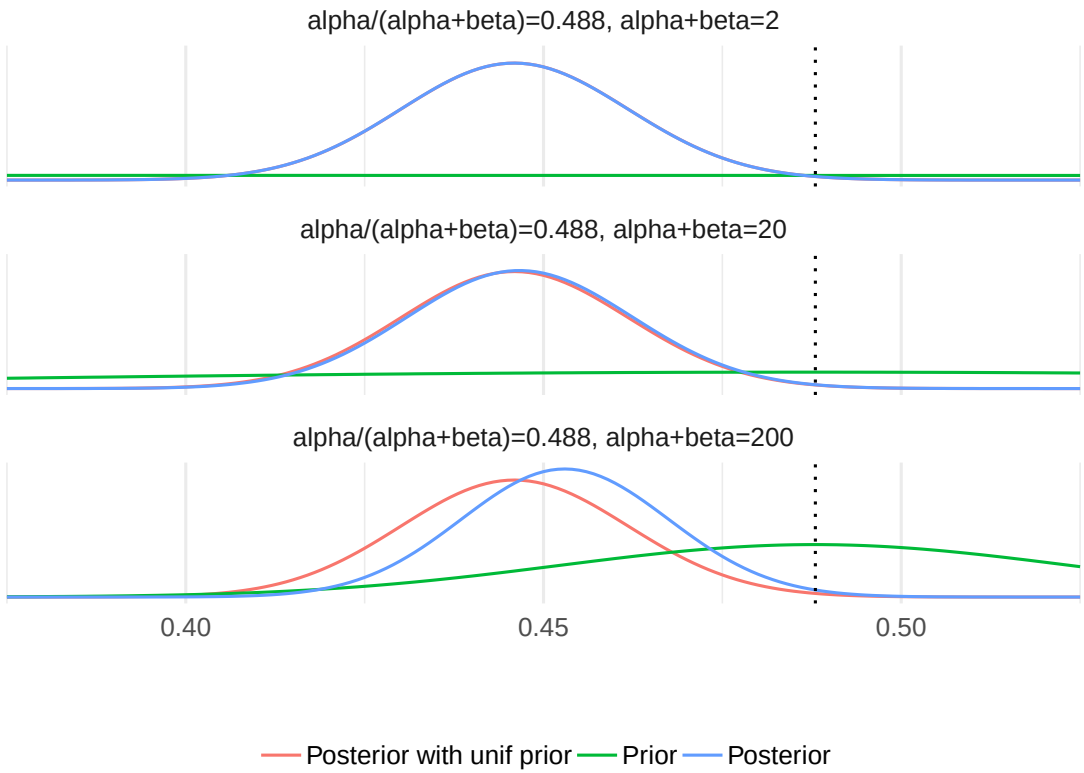
$$p(\theta|y,n,M) = \text{Beta}(\theta|\alpha+y,\beta+n-y)$$

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

UPPSALA UNIVERSITET

Placenta previa

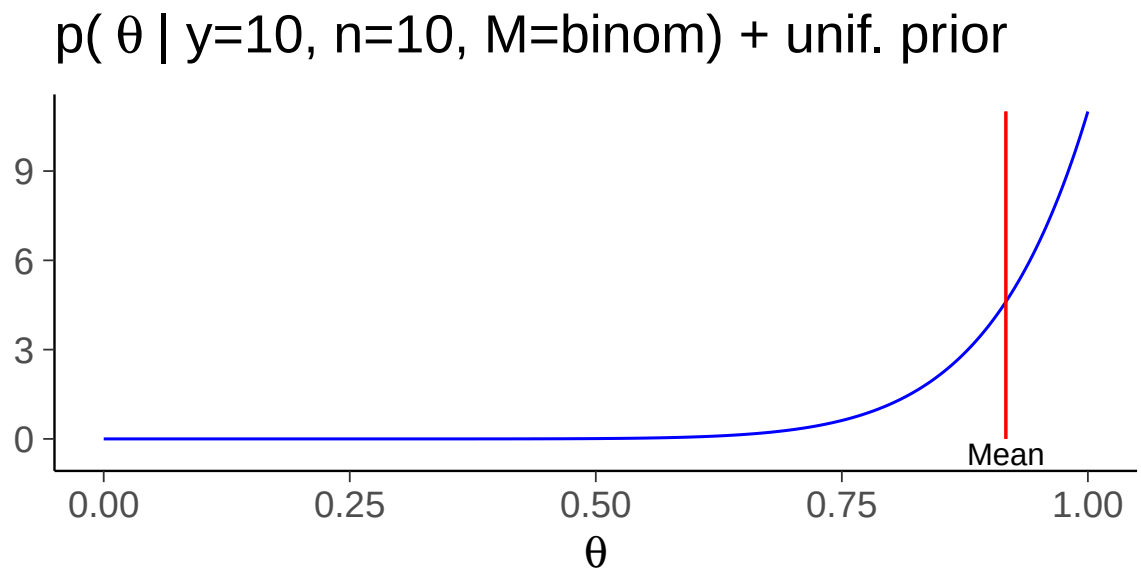
- Beta prior centered on population average 0.485

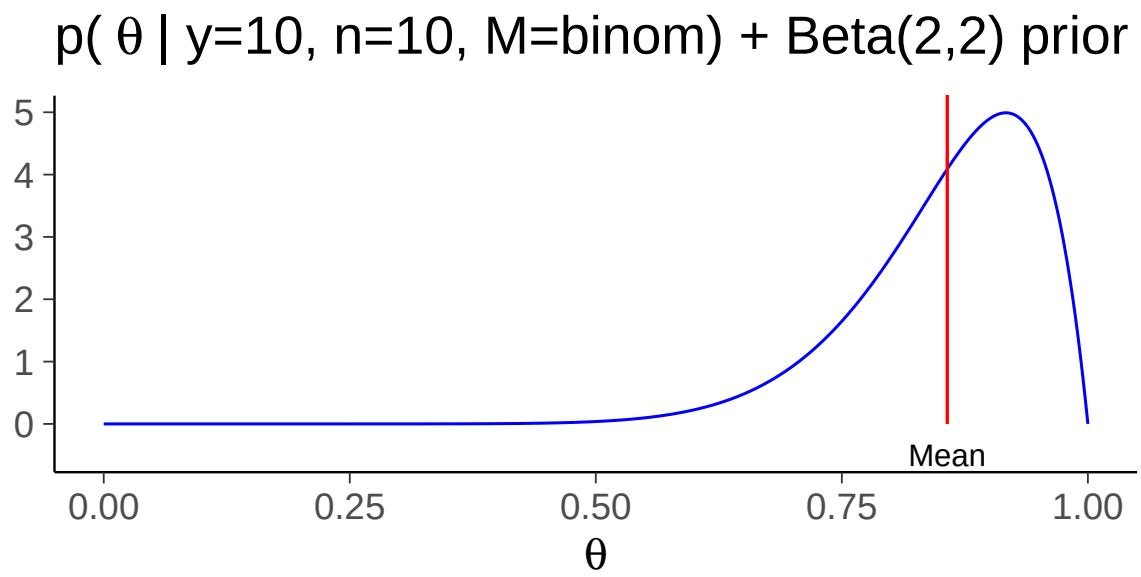


UPPSALA UNIVERSITET

Benefits of integration and prior

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





UPPSALA UNIVERSITET

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
  - when $n \to \infty$, $\text{E}[\theta|y] \to y/n$

- Posterior variance

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

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

UPPSALA UNIVERSITET

Noninformative prior

- Vague, flat, diffuse, or 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

UPPSALA UNIVERSITET

Proper and improper prior

- Proper prior has $\int p(\theta) = 1$
- Improper prior density doesn't have a finite integral
  - the posterior can still sometimes be proper
- Example: Binomial model
  - Beta(0,0) prior is improper
  - If $y \neq 0$ and $y \neq n$, the posterior is proper
- *Be careful with improper priors!*

UPPSALA UNIVERSITET

Weakly informative priors

- Weakly informative priors produce computationally better behaving posteriors
  - If we want to model IQ in children, how to construct a prior?
  - often there's some knowledge about the scale
  - Using the prior predictive distribution

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

    we can simulate data from the model:
    Does it look (remotely) reasonable?
  - useful if there's more information from previous observations - not certain how well that information is applicable in a new case

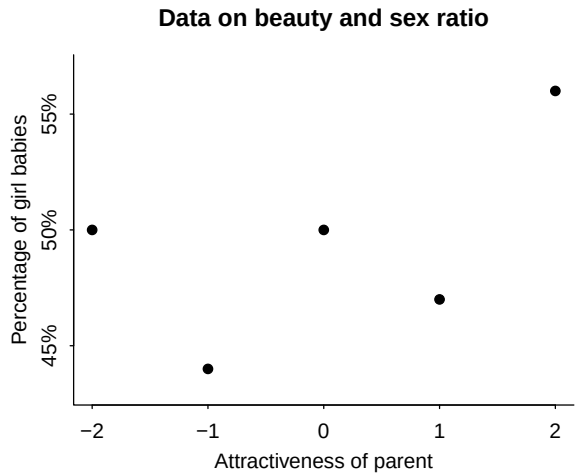
UPPSALA UNIVERSITET

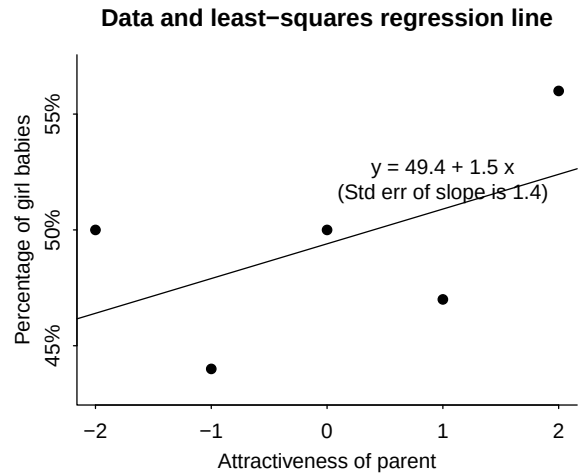
Construction of weakly informative priors

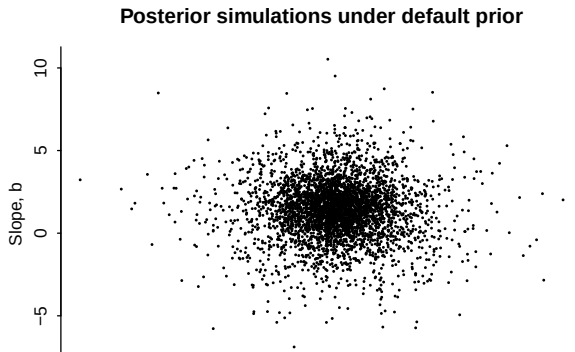
- Prior prediction checks!
- Start with some version of a noninformative prior, then add information until reasonable.
- Start with a strong prior, then broaden it to account for uncertainty
- Stan team prior choice recommendations
  https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

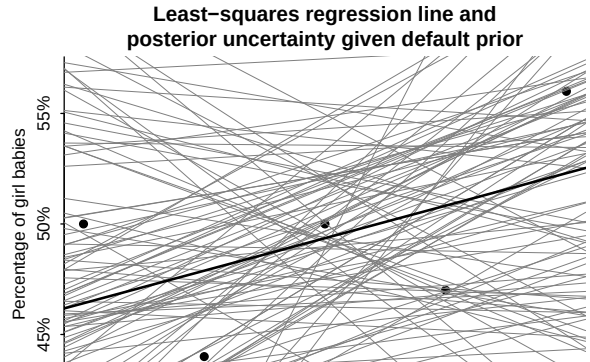
Example of informative prior

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





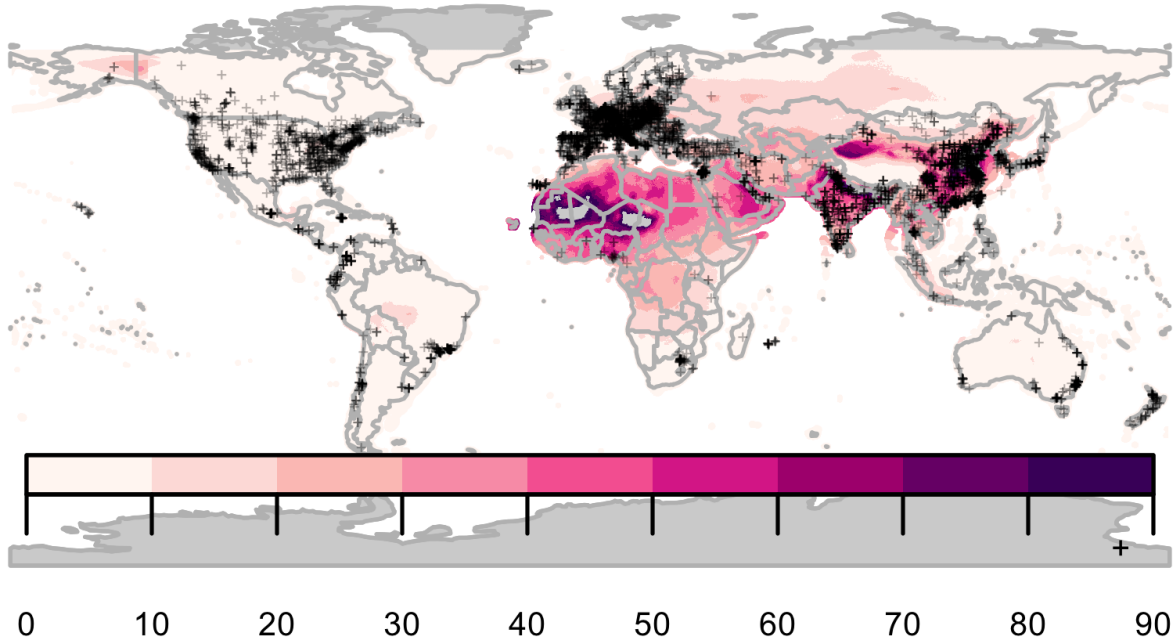




UPPSALA UNIVERSITET

Example of weakly informative prior

- Gabry et al (2019). 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)

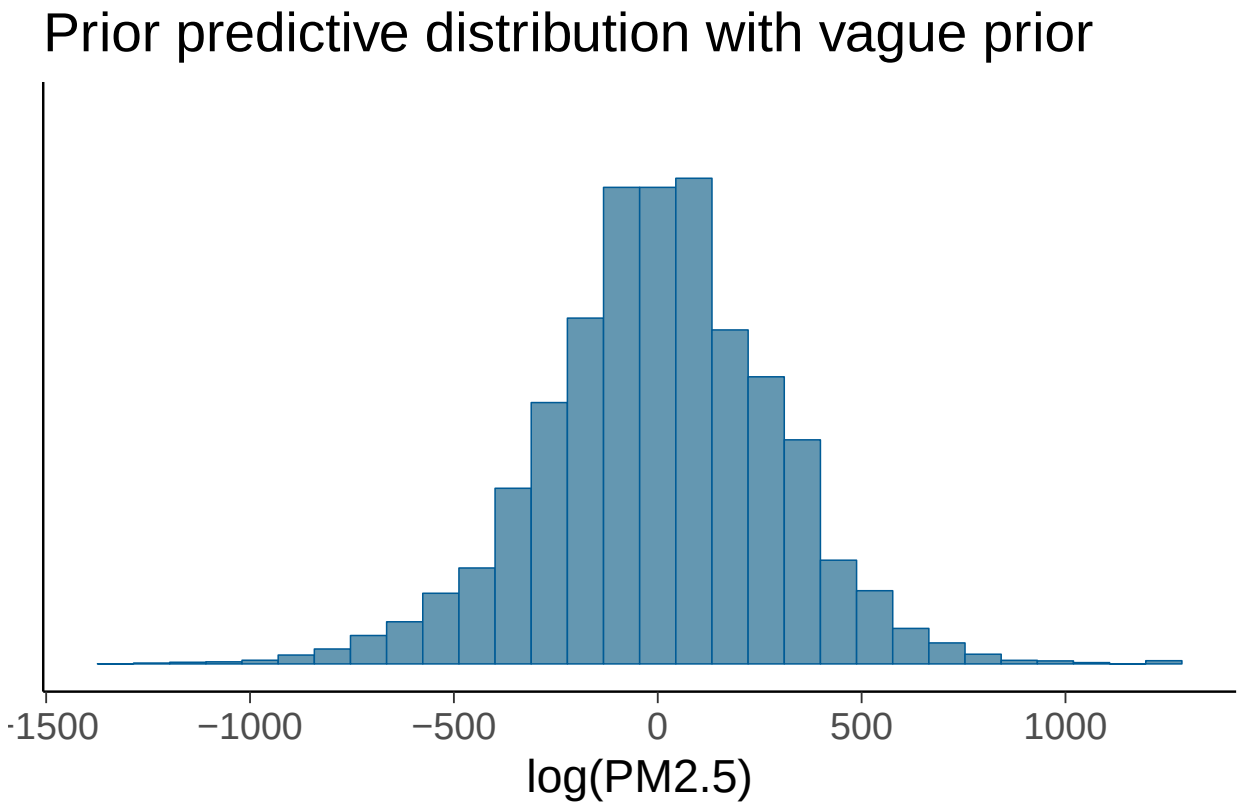


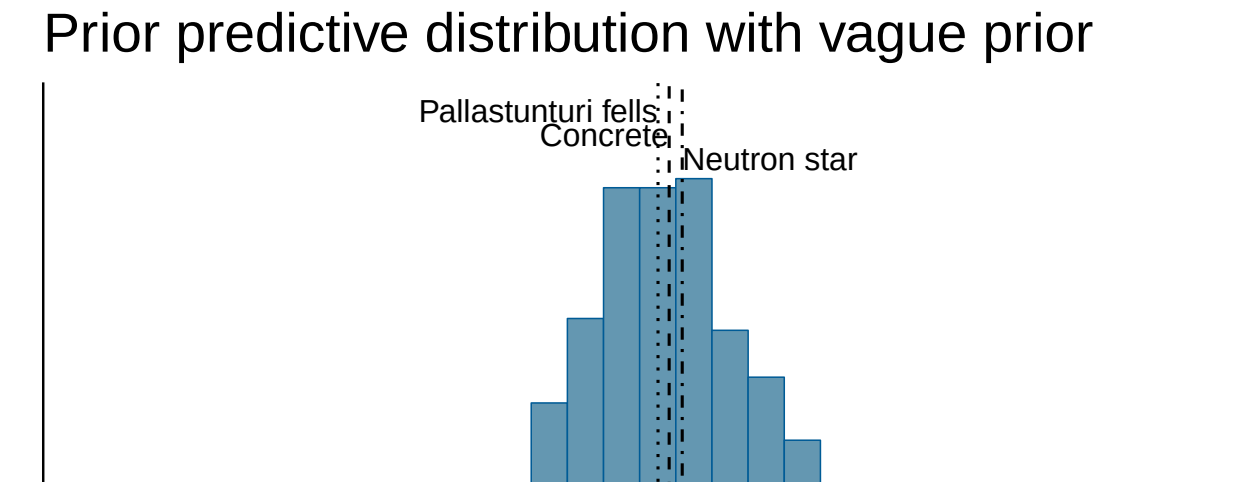
Satellite estimates and ground monitor locations

UPPSALA UNIVERSITET

- Posterior distributions
- Predictive distributions
- Prior distributions
- Demo
- The Normal model

Example of weakly informative prior





UPPSALA UNIVERSITET

Effect of incorrect priors?

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

UPPSALA UNIVERSITET

Sufficient statistics

- The function $t(y)$ of data y is said to be a *sufficient statistic* for θ if the likelihood for θ depends on the data y only through the value of $t(y)$.
- Example: Binomial model (with known n, and $y_i \in \{0, 1\}$)

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta) \prod^{n} p(y_i|\theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod^{n} \theta_i^{y}(1-\theta)^{1-y_i} \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \\
&\propto \theta^{\sum y_i+\alpha-1}(1-\theta)^{n-\sum y_i+\beta-1}
\end{aligned}
$$

after normalization

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

Hence, $\sum y_i$ is a sufficient statistic for θ in this model.

UPPSALA UNIVERSITET

Demo in R

- L2demo.R

UPPSALA UNIVERSITET

Algae

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.

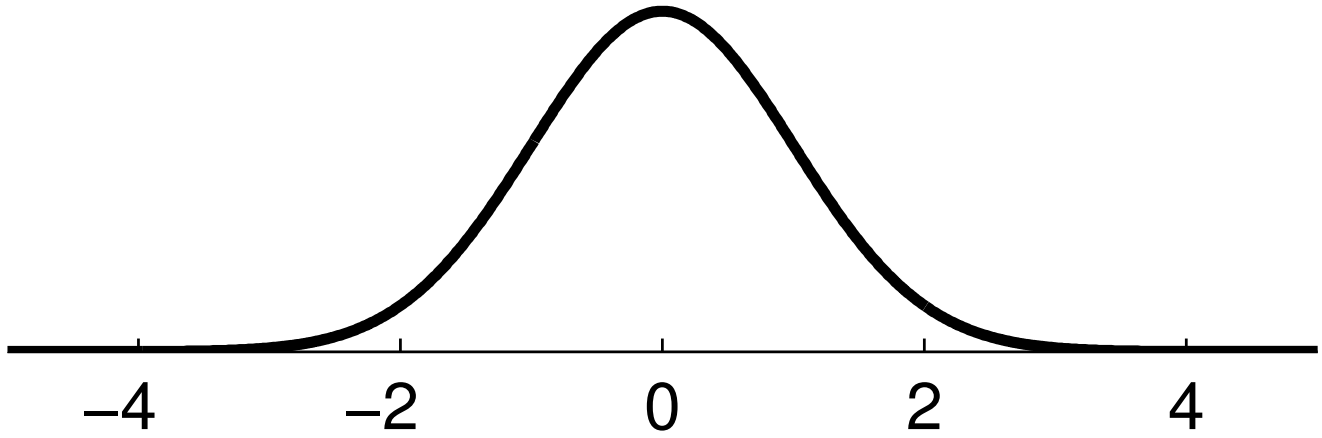
UPPSALA UNIVERSITET

Normal / Gaussian

- Observations $y \in \mathcal{R}$ (real valued)
- Mean θ and variance σ^2 (or deviation σ)
- For now: assume σ^2 is known

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

$$y \sim \mathcal{N}(\theta, \sigma^2)$$



UPPSALA UNIVERSITET

Reasons to use Normal distribution

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

UPPSALA UNIVERSITET

Central limit theorem (recap)

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

UPPSALA UNIVERSITET

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

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

UPPSALA UNIVERSITET

Normal distribution - conjugate prior for θ

- Posterior (see ex 2.14a)

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

$$\propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)$$

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

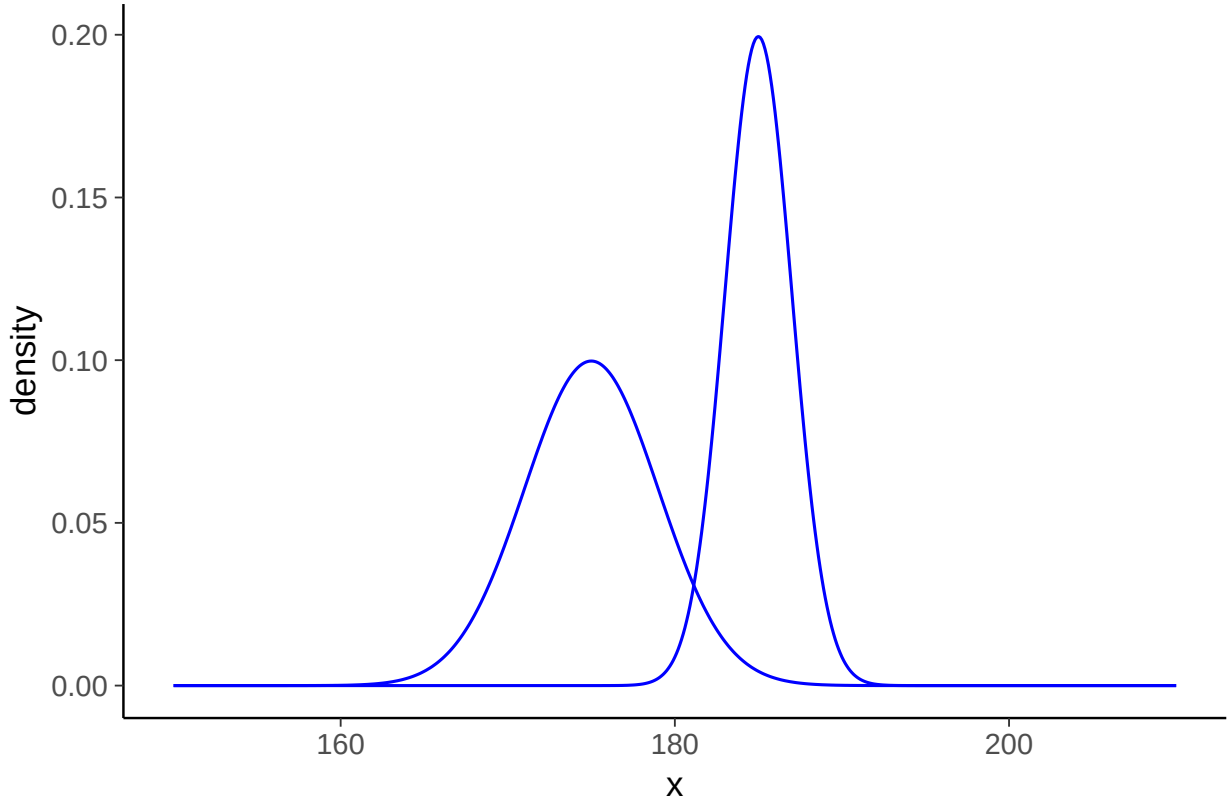
$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \text{and} \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

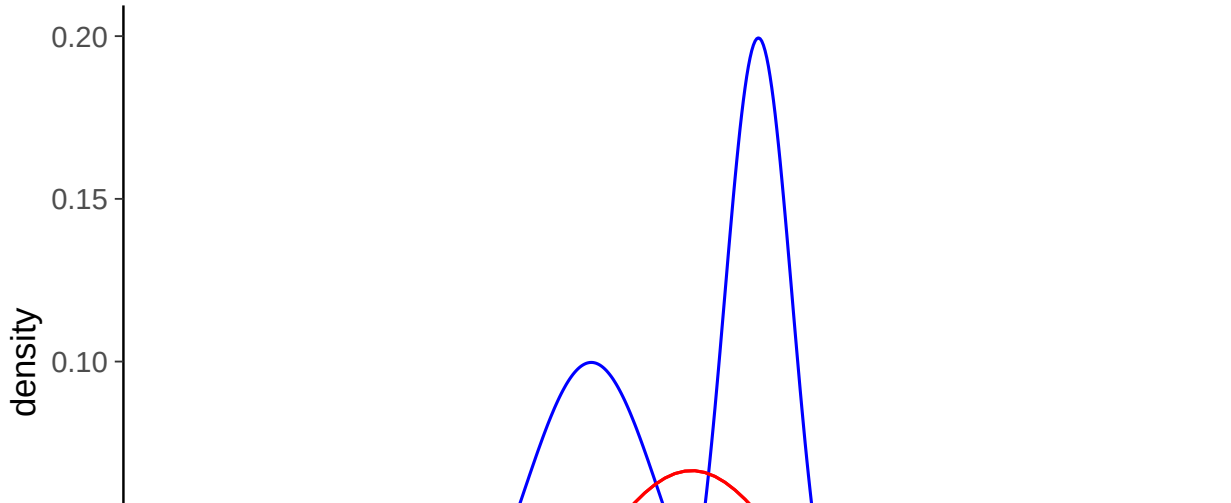
- 1/variance = precision
- Posterior precision = prior precision + data precision
- Posterior mean is precision weighted mean

UPPSALA UNIVERSITET

- Posterior distributions
- Predictive distributions
- Prior distributions
- Demo
- The Normal model

Normal distribution - example





UPPSALA UNIVERSITET

Normal distribution - conjugate prior for θ

Posterior (several observations $y = (y_1, \ldots, y_n)$)

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta)p(y|\theta) \\
&= p(\theta)\prod^{n} p(y_i|\theta) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{\sum^n (y_i-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right) \\
&= \exp\left(-\frac{1}{2}\left[\frac{n(\bar{y}-\theta)^2 + \sum^n (y_i-\bar{y})^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{n(\bar{y}-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)
\end{aligned}
$$

UPPSALA UNIVERSITET

Normal distribution - conjugate prior for θ

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

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

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

- 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

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

- Find the sufficient statistic for θ!

UPPSALA UNIVERSITET

Normal distribution - conjugate prior for θ

- Posterior predictive distribution

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

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

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

- Can be derived in multiple ways
  1. integrate
  2. $p(\tilde{y}, \theta)$ is a bivariate normal - marginalize out θ
- Predictive variance
  1. observation model variance σ^2
  2. posterior variance τ_1^2
- Aleatoric and epistemic uncertainty?