Outline of Lecture 2

- Binomial model is the simplest model
  - useful to introduce observation model, likelihood, posterior, prior, integration, posterior summaries
  - very commonly used as a building block
  - examples:
    - coin tossing
    - chips from bag
    - COVID tests and vaccines
    - classification / logistic regression

Outline of 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 unknown 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)
- 2.8–2.9 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 several events in independent trials is e.g. $\theta\theta(1-\theta)\theta(1-\theta)(1-\theta)\ldots$

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

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

Binomial: known θ

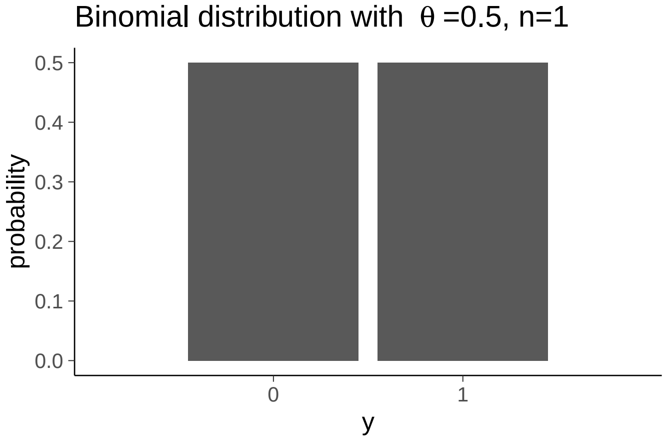
- Observation model (function of y, discrete)

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

Binomial: known θ

- Observation model (function of y, discrete)

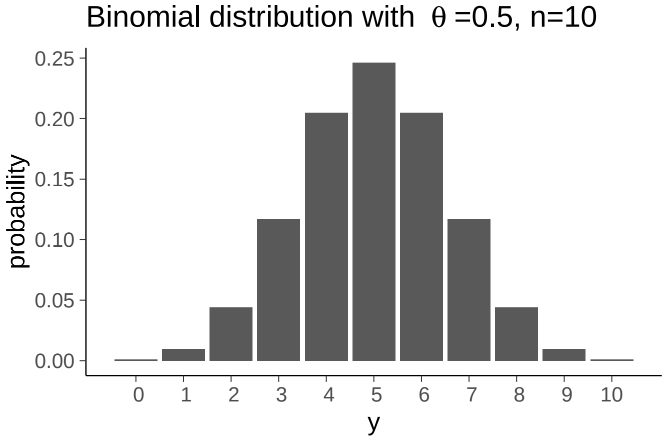
$$p(y|\theta, n) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$



Binomial: known θ

- Observation model (function of y, discrete)

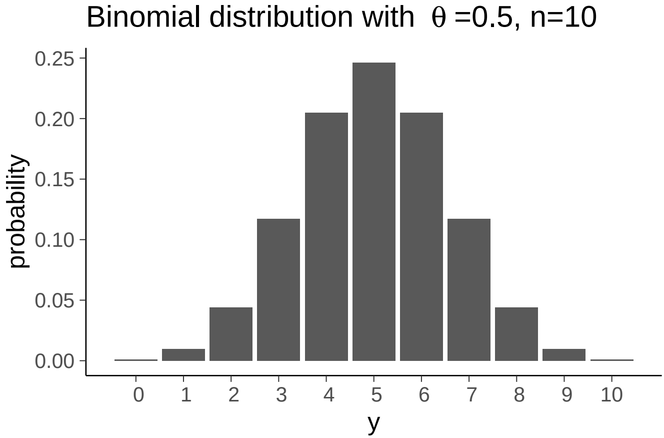
$$p(y|\theta, n) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$



Binomial: known θ

- Observation model (function of y, discrete)

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

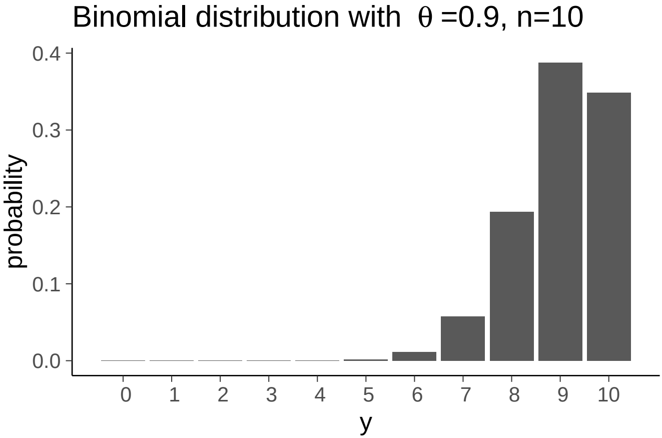


$p(y|n = 10, \theta = 0.5)$: 0.00 0.01 0.04 0.12 0.21 0.25 0.21 0.12 0.04 0.01 0.00

Binomial: known θ

- Observation model (function of y, discrete)

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

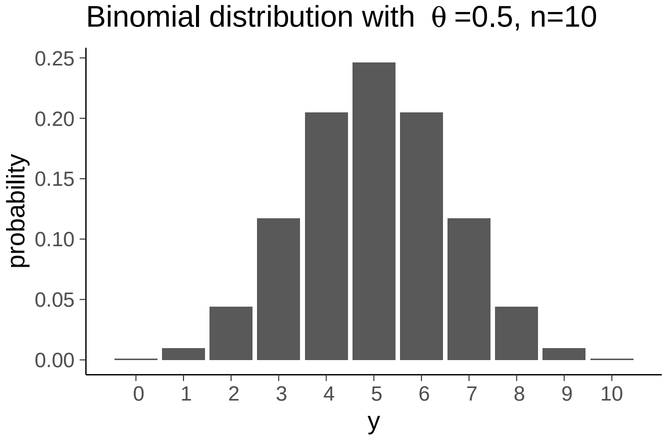


$p(y|n=10, \theta=0.9)$: 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.06 0.19 0.39 0.35

Binomial: what if $y = 6$?

- Observation model (function of y, discrete)

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

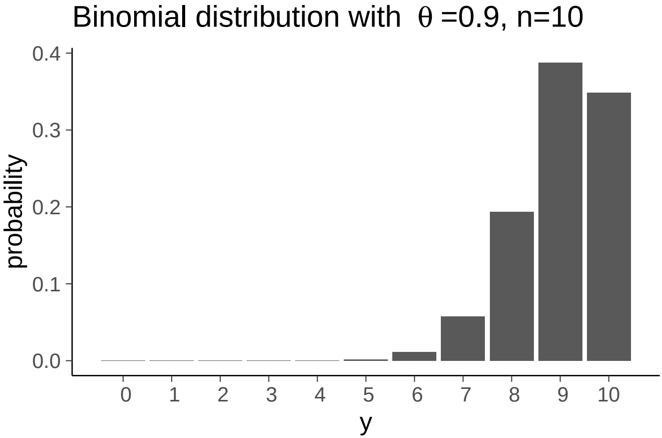


$p(y = 6|n = 10, \theta = 0.5)$: 0.00 0.01 0.04 0.12 0.21 0.25 **0.21** 0.12 0.04 0.01 0.00

Binomial: what if $y = 6$?

- Observation model (function of y, discrete)

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

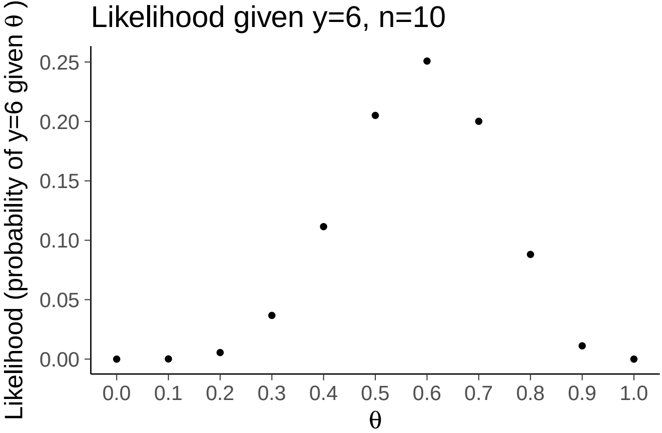


$p(y = 6|n = 10, \theta = 0.9)$: 0.00 0.00 0.00 0.00 0.00 0.00 **0.01** 0.06 0.19 0.39 0.35

Binomial: unknown θ and $y = 6$

- Likelihood (function of θ, continuous)

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

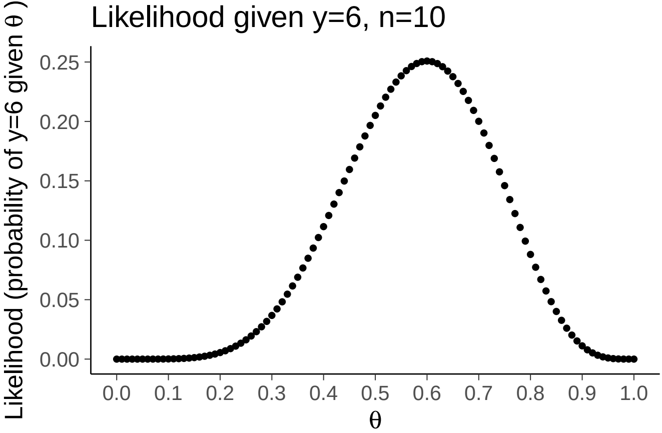


$p(y = 6|n = 10, \theta)$: 0.00 0.00 0.01 0.04 0.11 **0.21** 0.25 0.20 0.09 **0.01** 0.00

Binomial: unknown θ and $y = 6$

- Likelihood (function of θ, continuous)

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

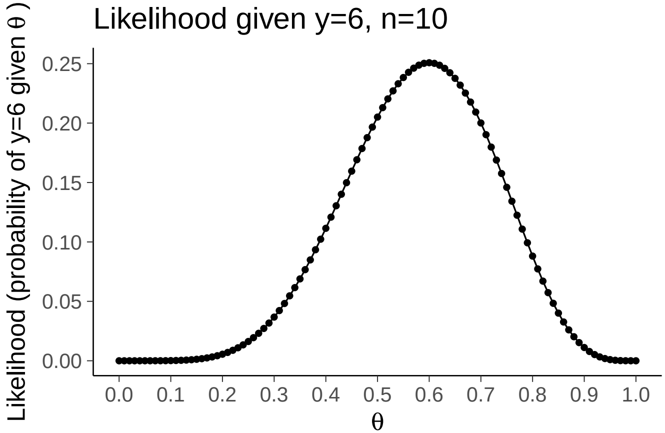


we can compute the value for any θ, but in practice can compute only for finite values

Binomial: unknown θ and $y = 6$

- Likelihood (function of θ, continuous)

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

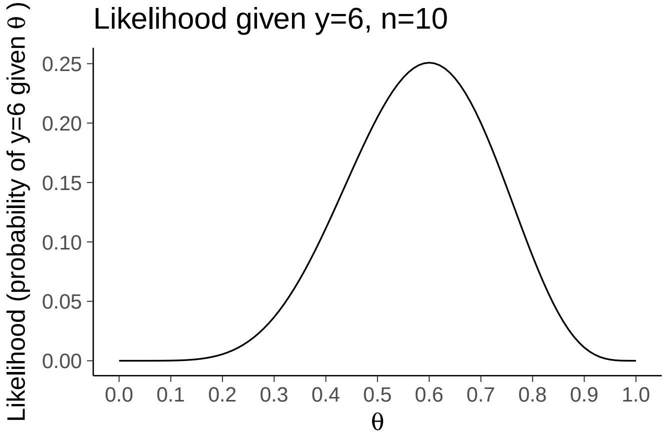


with sufficient many evaluations, linearly interpolated plot looks smooth

Binomial: unknown θ and $y = 6$

- Likelihood (function of θ, continuous)

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

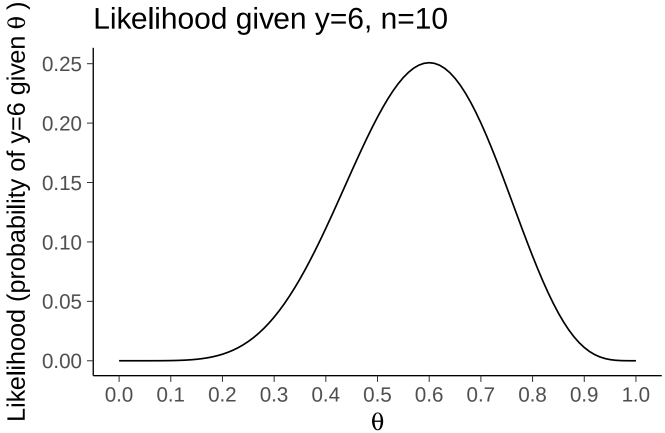


looks smooth, and we'll get back to later to computational cost issues

Binomial: unknown θ and $y = 6$

- Likelihood (function of θ, continuous)

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

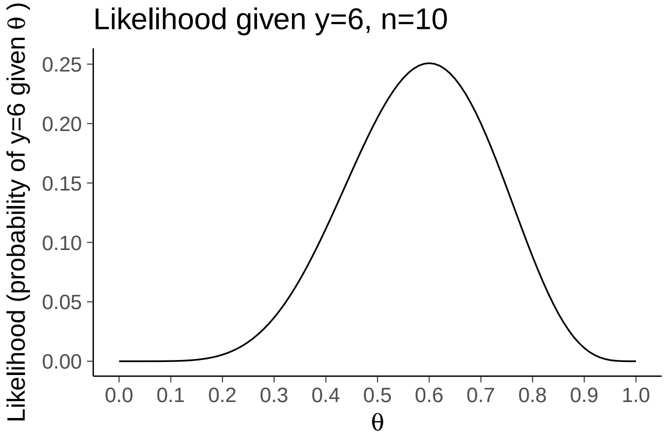


likelihood function describes uncertainty, but is not normalized distribution

Binomial: unknown θ and $y = 6$

- Likelihood (function of θ, continuous)

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



`integrate(function(theta) dbinom(6, 10, theta), 0, 1)` $\approx 0.09 \neq 1$

Binomial posterior

- Joint distribution $p(\theta, y|n)$
  - Observation model as a function of y: $p(y|\theta, n) \propto p(\theta, y|n)$
  - Likelihood as a function of θ: $p(y|\theta, n) \propto p(\theta, y|n)$

Binomial posterior

- Joint distribution $p(\theta, y|n)$
  - Observation model as a function of y: $p(y|\theta, n) \propto p(\theta, y|n)$
  - Likelihood as a function of θ: $p(y|\theta, n) \propto p(\theta, y|n)$
- Posterior with Bayes rule (function of θ, continuous)

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

Binomial posterior

- Joint distribution $p(\theta, y|n)$
  - Observation model as a function of y: $p(y|\theta, n) \propto p(\theta, y|n)$
  - Likelihood as a function of θ: $p(y|\theta, n) \propto p(\theta, y|n)$
- Posterior with Bayes rule (function of θ, continuous)

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

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

Binomial posterior

- Joint distribution $p(\theta, y|n)$
  - Observation model as a function of y: $p(y|\theta, n) \propto p(\theta, y|n)$
  - Likelihood as a function of θ: $p(y|\theta, n) \propto p(\theta, y|n)$
- Posterior with Bayes rule (function of θ, continuous)

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

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

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

Binomial posterior

- Joint distribution $p(\theta, y|n)$
  - Observation model as a function of y: $p(y|\theta, n) \propto p(\theta, y|n)$
  - Likelihood as a function of θ: $p(y|\theta, n) \propto p(\theta, y|n)$
- Posterior with Bayes rule (function of θ, continuous)

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

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

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

- Then

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

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

Binomial: unknown θ

- Normalization term Z (constant given y)

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

Binomial: unknown θ

- Normalization term Z (constant given y)

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

- Evaluate with $y = 6, n = 10$

`y<-6;n<-10;`

`integrate(function(theta) theta^y*(1-theta)^(n-y), 0, 1)` ≈ 0.0004329

`gamma(6+1)*gamma(10-6+1)/gamma(10+2)` ≈ 0.0004329

Binomial: unknown θ

- Normalization term Z (constant given y)

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

- Evaluate with $y = 6, n = 10$

```
y<-6;n<-10;
```

`integrate(function(theta) theta^y*(1-theta)^(n-y), 0, 1)` ≈ 0.0004329

`gamma(6+1)*gamma(10-6+1)/gamma(10+2)` ≈ 0.0004329
usually computed via $\log\Gamma(\cdot)$ due to the limitations of floating point representation

Binomial: unknown θ

- Normalization term Z (constant given y)

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

- Evaluate with $y = 6, n = 10$

`y<-6;n<-10;`

`integrate(function(theta) theta^y*(1-theta)^(n-y), 0, 1)` ≈ 0.0004329

`gamma(6+1)*gamma(10-6+1)/gamma(10+2)` ≈ 0.0004329
usually computed via $\log\Gamma(\cdot)$ due to the limitations of floating point representation

- Normalization term in closed form is not in general available
    - later you learn approximate integration methods that work without knowing the normalization term

Binomial: unknown θ

- Posterior is

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

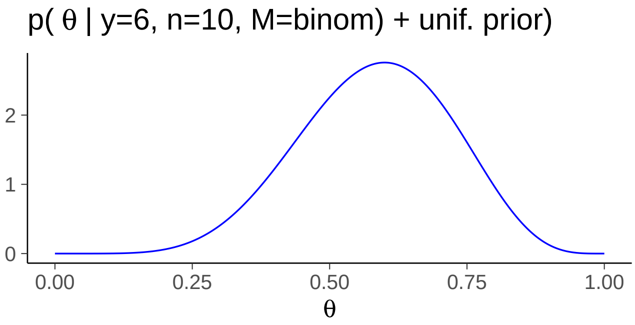
Binomial: unknown θ

- Posterior is

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

which is called Beta distribution

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



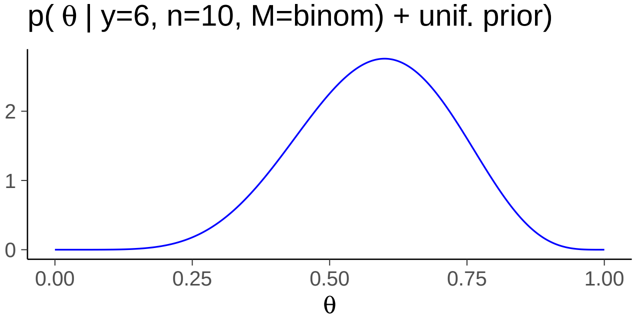
Binomial: computation

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

  - `from preliz import Beta`
  - Similar to scipy.stats
  - Nicer features to explore distributions
  - https://preliz.readthedocs.io

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, BDA3 Ch 4), because he didn’t know how to compute Beta CDF

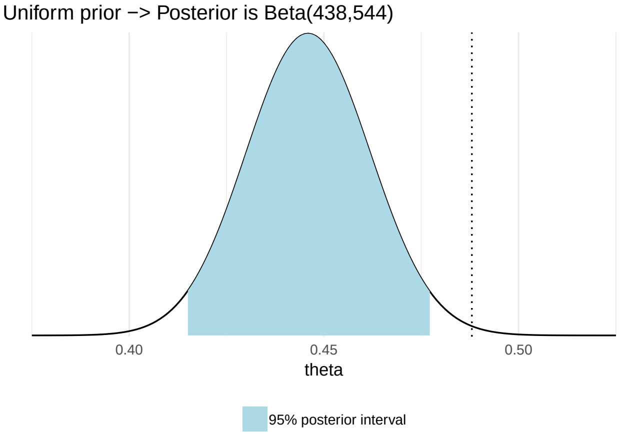


Placenta previa

- Probability of a girl birth given placenta previa (BDA3 p. 37)
  - 437 girls and 543 boys have been observed
  - is the proportion of girls 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 proportion of girls 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

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

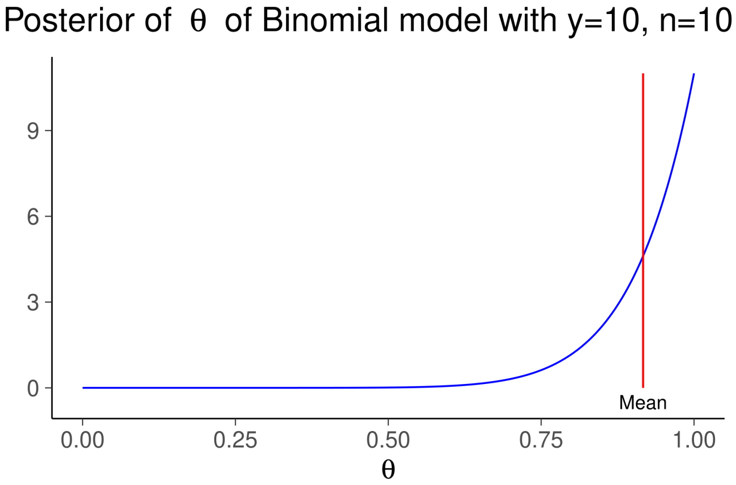
- Extreme cases

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

- cf. maximum likelihood

Benefits of integration

Example: $n = 10, y = 10$



Predictive distribution

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

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

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

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

Left handedness

- If we would like to provide scissors for all students, how many left handed scissors we would need?

Left handedness

- If we would like to provide scissors for all students, how many left handed scissors we would need?
- Posterior distribution for θ is $\mathrm{Beta}(\alpha + l, \beta + r)$

Left handedness

- If we would like to provide scissors for all students, how many left handed scissors we would need?
- Posterior distribution for θ is $\mathrm{Beta}(\alpha + l, \beta + r)$
- Posterior predictive distribution for $\tilde{l}$ is
  $\text{Beta-Binomial}(\tilde{l}|N - n, \alpha + l, \beta + r) =$
  $\int_0^1 \mathrm{Bin}(\tilde{l}|N - n, \theta)\, \mathrm{Beta}(\theta|\alpha + l, \beta + r) d\theta$

Left handedness

- If we would like to provide scissors for all students, how many left handed scissors we would need?
- Posterior distribution for θ is $\mathrm{Beta}(\alpha + l, \beta + r)$
- Posterior predictive distribution for $\tilde{l}$ is $\mathrm{Beta\text{-}Binomial}(\tilde{l}|N - n, \alpha + l, \beta + r) = \int_0^1 \mathrm{Bin}(\tilde{l}|N - n, \theta)\,\mathrm{Beta}(\theta|\alpha + l, \beta + r) d\theta$
- Demo: https://handedness.streamlit.app

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 / NUTS used e.g. in Stan no 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

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

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

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

Beta prior for Binomial model

- Prior

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

- Posterior

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

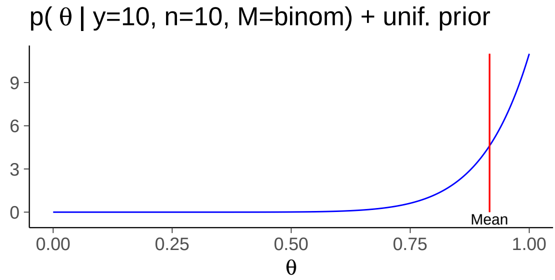
after normalization

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

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

Benefits of integration and prior

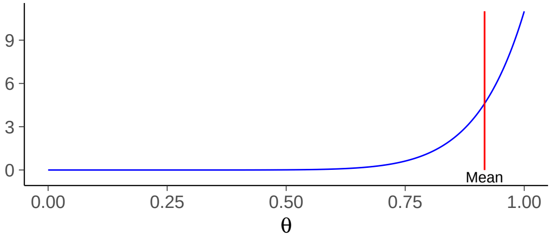
Example: $n = 10, y = 10$ - uniform vs Beta(2,2) prior



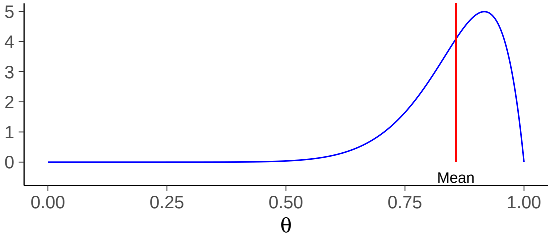
Benefits of integration and prior

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

p(θ | y=10, n=10, M=binom) + unif. prior



p(θ | y=10, n=10, M=binom) + 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

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

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

Beta prior for Binomial model

- Posterior

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

- Posterior mean

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

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

- 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 improper 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
- 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

Informative prior for left handedness

- Papadatou-Pastou et al. (2020). Human handedness: A meta-analysis. Psychological Bulletin, 146(6), 481–524. https://doi.org/10.1037/bul0000229
  - totaling 2 396 170 individuals
  - varies between 9.3% and 18.1%, depending on how handedness is measured
  - varies between countries and in time

Informative prior for left handedness

- Papadatou-Pastou et al. (2020). Human handedness: A meta-analysis. Psychological Bulletin, 146(6), 481–524. https://doi.org/10.1037/bul0000229
  - totaling 2 396 170 individuals
  - varies between 9.3% and 18.1%, depending on how handedness is measured
  - varies between countries and in time

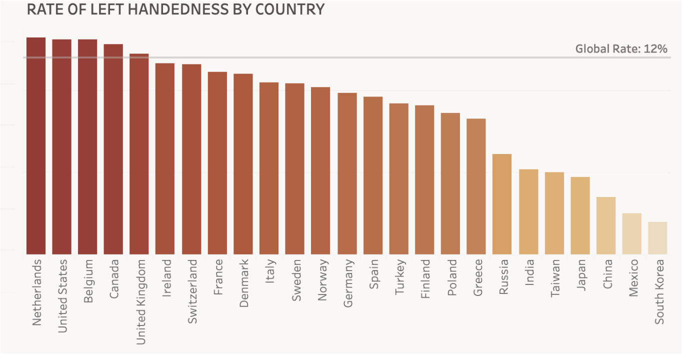


Fig from https://www.reddit.com/r/dataisbeautiful/comments/s9x1ya/history_of_lefthandedness_oc/

Informative prior for left handedness

- Papadatou-Pastou et al. (2020). Human handedness: A meta-analysis. Psychological Bulletin, 146(6), 481–524. https://doi.org/10.1037/bul0000229
  - totaling 2 396 170 individuals
  - varies between 9.3% and 18.1%, depending on how handedness is measured
  - varies between countries and in time

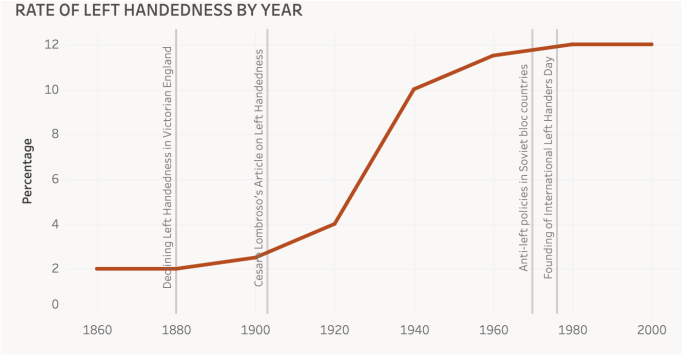
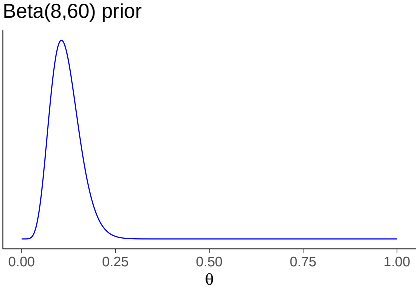


Fig from https://www.reddit.com/r/dataisbeautiful/comments/s9x1ya/history_of_lefthandedness_oc/

Informative prior for left handedness

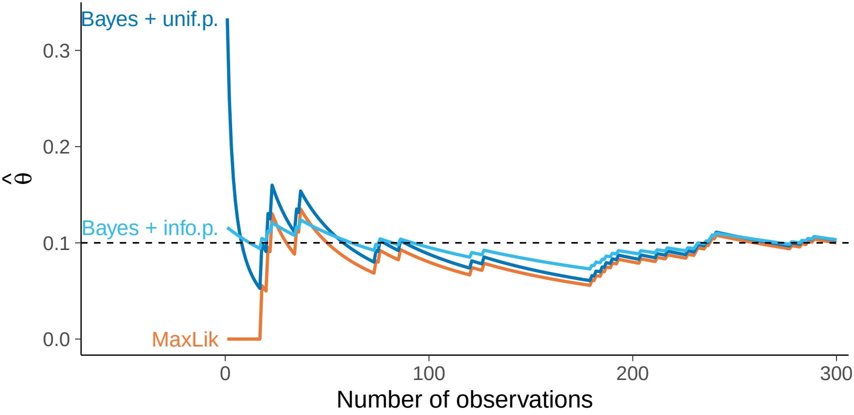
- Papadatou-Pastou et al. (2020). Human handedness: A meta-analysis. Psychological Bulletin, 146(6), 481–524. https://doi.org/10.1037/bul0000229
  - totaling 2 396 170 individuals
  - varies between 9.3% and 18.1%, depending on how handedness is measured
  - varies between countries and in time



Demo:https://handedness.streamlit.app

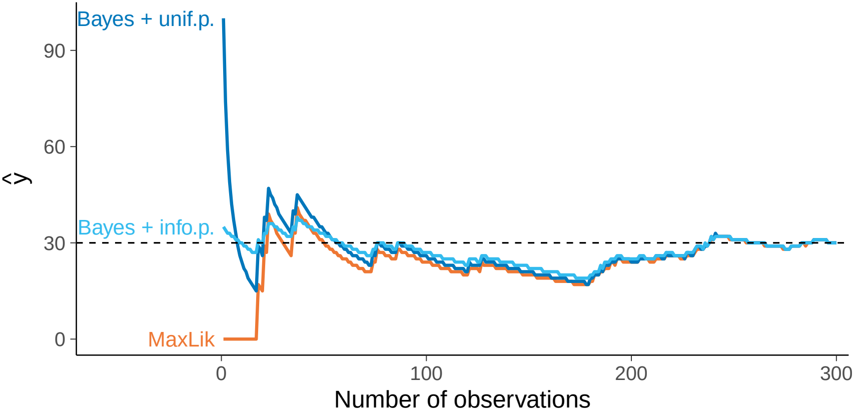
Benefits of integration and prior

- Left handed simulation with $L = 30$ left handed and $N = 300$ total



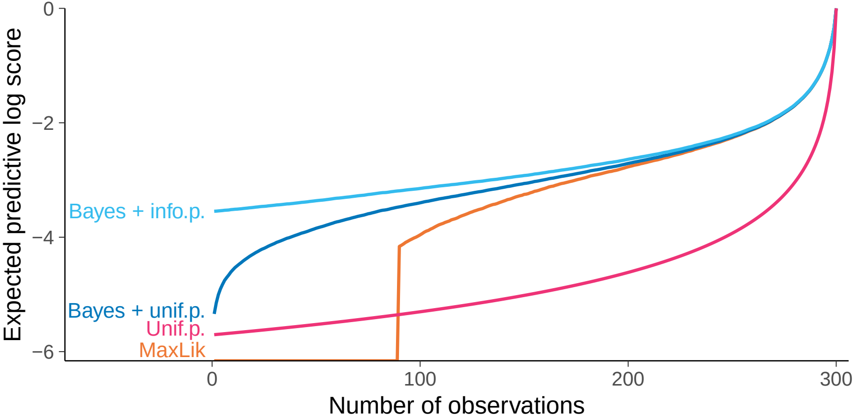
Benefits of integration and prior

- Left handed simulation with $L = 30$ left handed and $N = 300$ total



Benefits of integration and prior

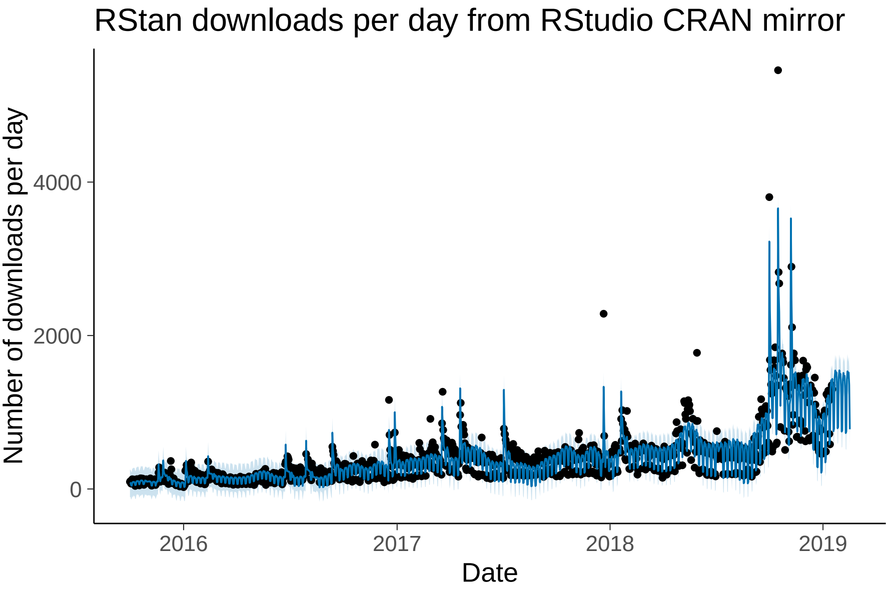
- Left handed simulation with true $\theta = 0.1$ and $N = 300$
  - repeated 10 000 times
  - average log predictive probability for guessing L after $n \le N$ observations



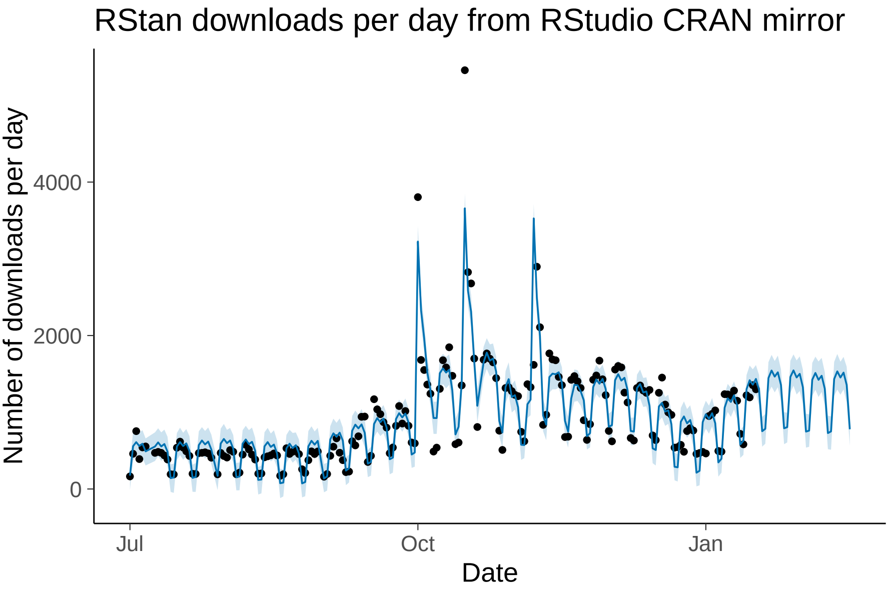
Effect of incorrect priors?

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

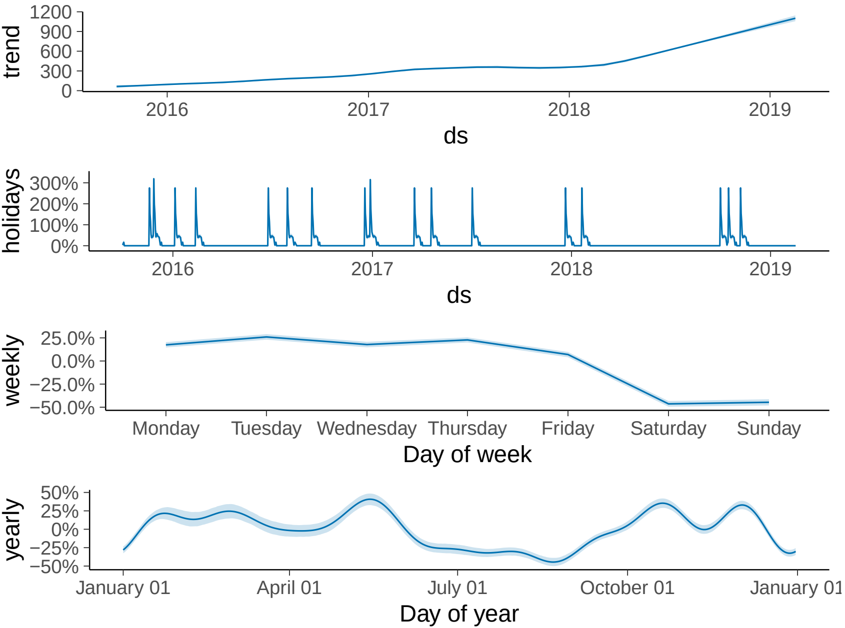
Structural information in predicting future



Structural information in predicting future



Structural information – Prophet by Facebook



Binomial: unknown θ

Sometimes conditioning on the model M is explicitly shown

- 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

Sometimes conditioning on the model M is explicitly shown

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

- makes it more clear that likelihood and prior are both part of the model
- makes it more clear that there is no absolute probability for $p(y|n)$, but it depends on the model M
- in case of two models, we can evaluate marginal likelihoods $p(y|n,M_1)$ and $p(y|n,M_2)$ (more in Ch 7)

Binomial: unknown θ

Sometimes conditioning on the model M is explicitly shown

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

- makes it more clear that likelihood and prior are both part of the model
- makes it more clear that there is no absolute probability for $p(y|n)$, but it depends on the model M
- in case of two models, we can evaluate marginal likelihoods $p(y|n,M_1)$ and $p(y|n,M_2)$ (more in Ch 7)
- usually dropped to make the notation more concise

Sufficient statistics

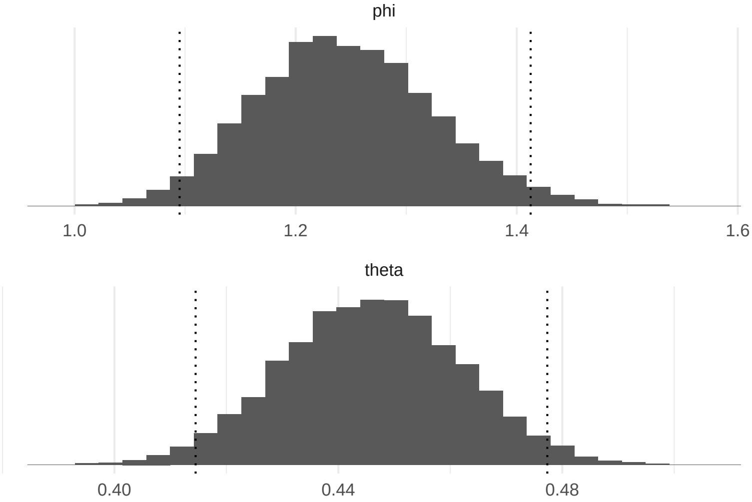
- The quantity $t(y)$ is said to be a *sufficient statistic* for θ, because the likelihood for θ depends on the data y only through the value of $t(y)$.

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 $t(y)$.
- For binomial model the sufficient statistics are y and n (the order doesn't matter)

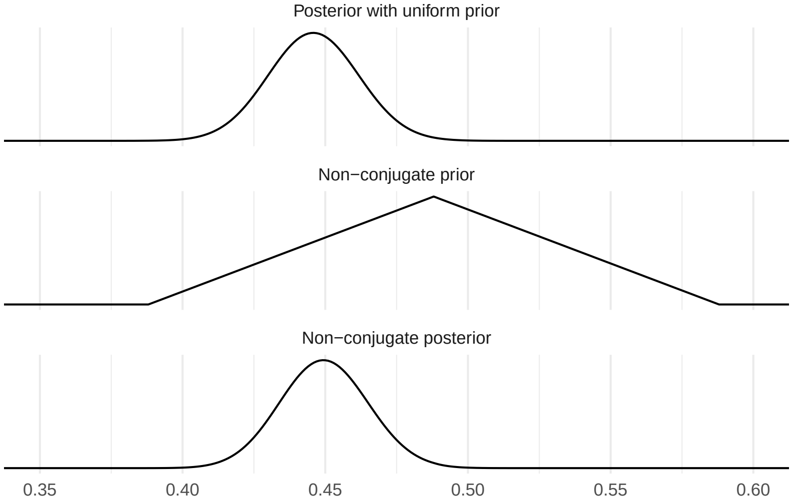
Posterior visualization and inference demos

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



Posterior visualization and inference demos

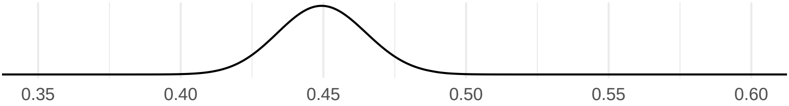
- demo2_4: Compute posterior distribution in a grid.



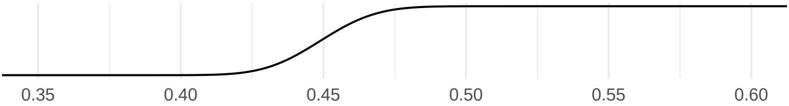
Posterior visualization and inference demos

- demo2_4: Sample using the inverse-cdf method.

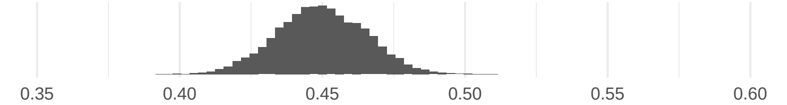
Non−conjugate posterior



Posterior−cdf



Histogram of posterior samples



Algae

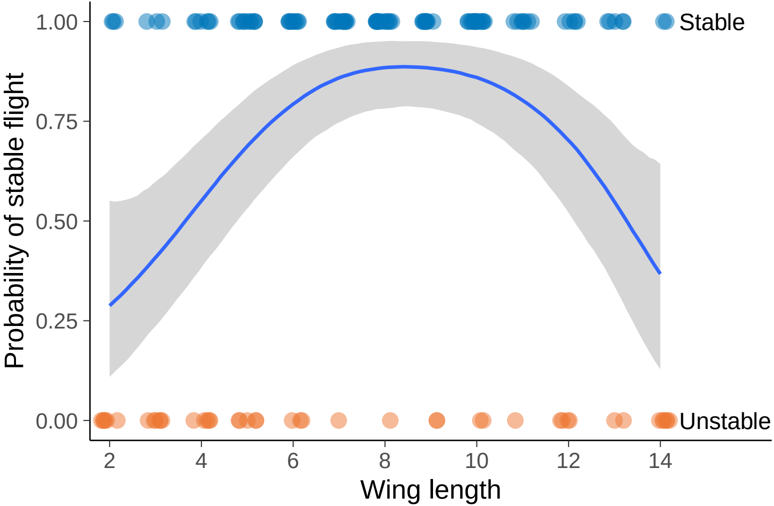
Assignment

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.(rda|txt)* ('0': no algae, '1': algae present). Let θ be the probability of a monitoring site having detectable blue-green algae levels.

- Use a binomial model for observations and a $\mathrm{Beta}(2, 10)$ prior.
- What can you say about the value of the unknown θ?
- Experiment how the result changes if you change the prior.

Binomial model with $\theta = f(x)$

- Next week you learn how the binomial model parameter θ can depend on some other measurement x

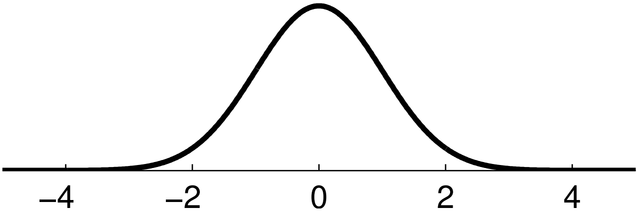


Normal / Gaussian

- Observations y real valued
- Mean θ and variance σ^2 (or deviation σ)
  This week assume σ^2 known (preparing for the next week)

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

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



Reasons to use Normal distribution

- Computational convenience
- Tradition
- Often good approximation (justification based on central limit theorem)

Central limit theorem*

- Given certain conditions, distribution of sum (and mean) of random variables approach Gaussian distribution as $n \to \infty$

Central limit theorem*

- Given certain conditions, distribution of sum (and mean) of random variables approach Gaussian distribution as $n \to \infty$
- Problems
  - does not hold for distributions with infinite variance, e.g., Cauchy

Central limit theorem*

- Given certain conditions, distribution of sum (and mean) of random variables approach Gaussian distribution as $n \to \infty$
- Problems
  - does not hold for distributions with infinite variance, e.g., Cauchy
  - may require large n,
    e.g. Binomial, when θ close to 0 or 1

Central limit theorem*

- Given certain conditions, distribution of sum (and mean) of random variables approach Gaussian distribution as $n \to \infty$
- Problems
  - does not hold for distributions with infinite variance, 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 (highly recommended to do BDA 3 Ex 2.14a)

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

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

Normal distribution - conjugate prior for θ

- Posterior (highly recommended to do BDA 3 Ex 2.14a)

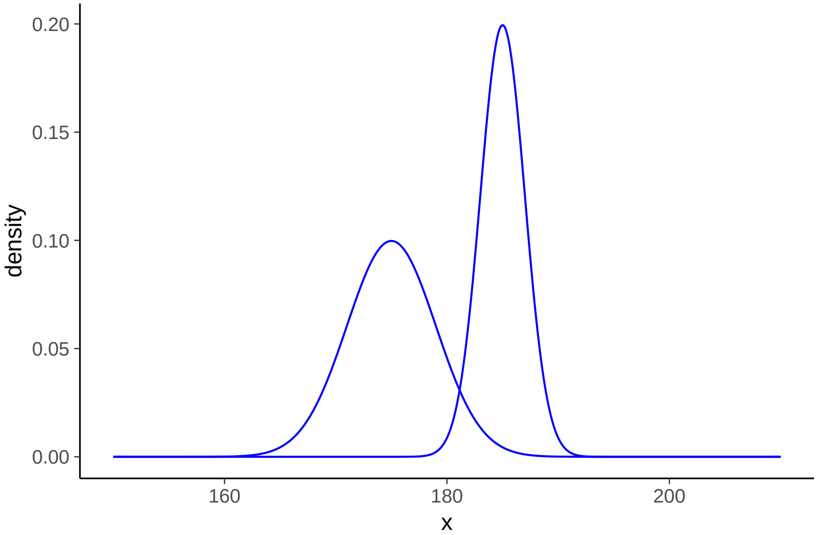
$$p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right]\right)$$

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

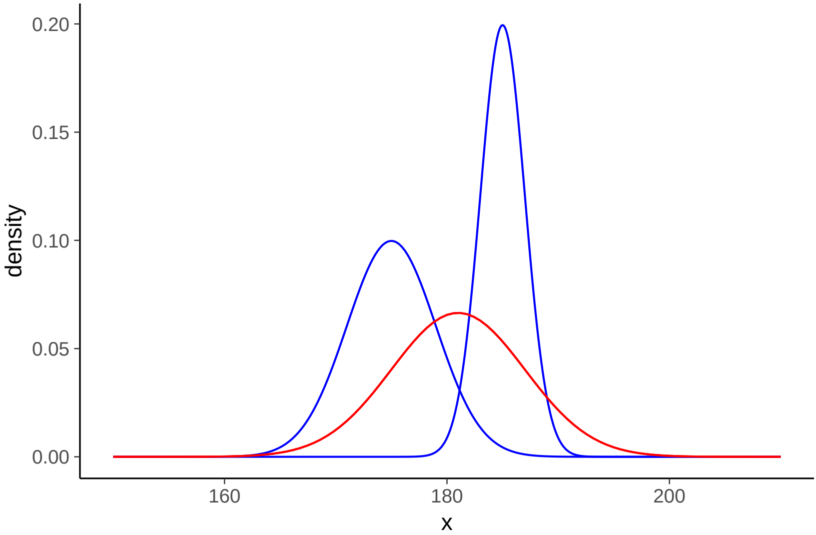
$\theta|y \sim \mathrm{N}(\mu_1, \tau_1^2)$, where $\mu_1 = \dfrac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$ and $\dfrac{1}{\tau_1^2} = \dfrac{1}{\tau_0^2} + \dfrac{1}{\sigma^2}$

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

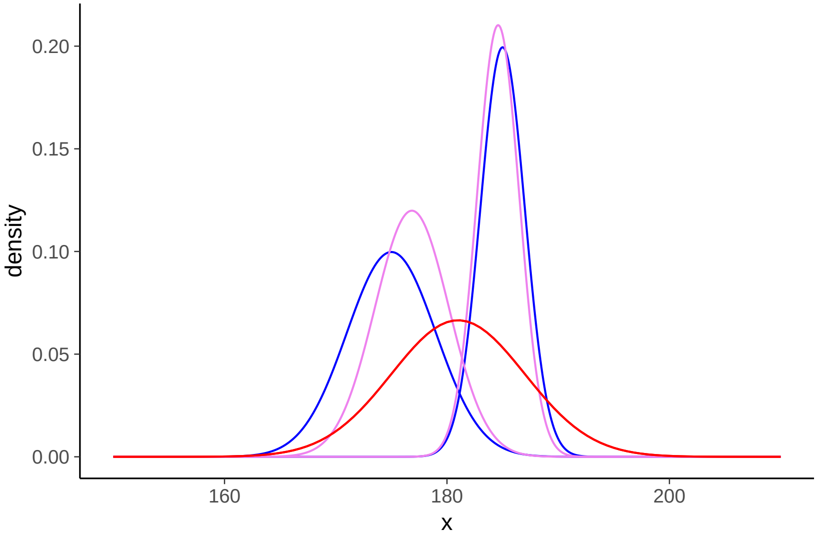
Normal distribution - example



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 $y = (y_1, \ldots, y_n)$

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

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

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

Normal distribution - conjugate prior for θ

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

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

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

- 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

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

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

Normal model

- Gets more interesting when both mean and variance are unknown
  - next week

Normal model

- Gets more interesting when both mean and variance are unknown
  - next week
- The mean can be also a function of covariates
  - e.g. normal linear regression $y \sim \mathrm{N}(\alpha + \beta x, \sigma^2)$

Normal model

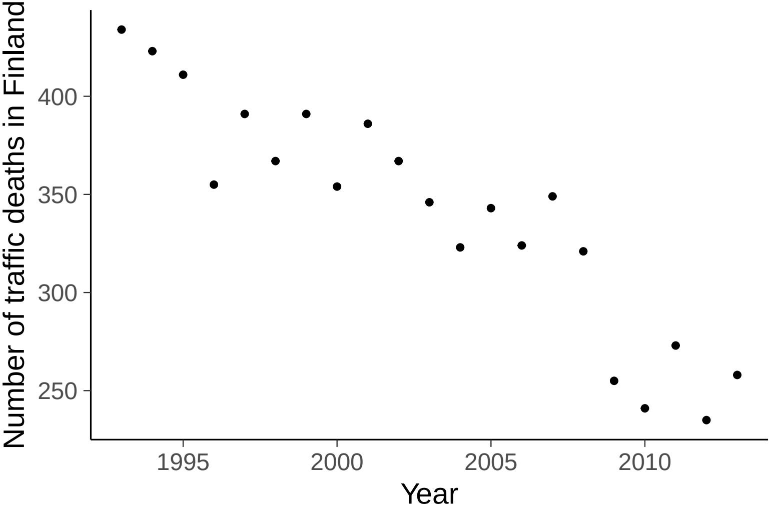
- Gets more interesting when both mean and variance are unknown
  - next week
- The mean can be also a function of covariates
  - e.g. normal linear regression $y \sim \mathrm{N}(\alpha + \beta x, \sigma^2)$
- Gaussian processes, Kalman filters, variational inference, Laplace approximaion, etc.

Some other one parameter models

- Poisson, useful for count data (e.g. in epidemiology)
- Exponential, useful for time to an event (e.g. particle decay)

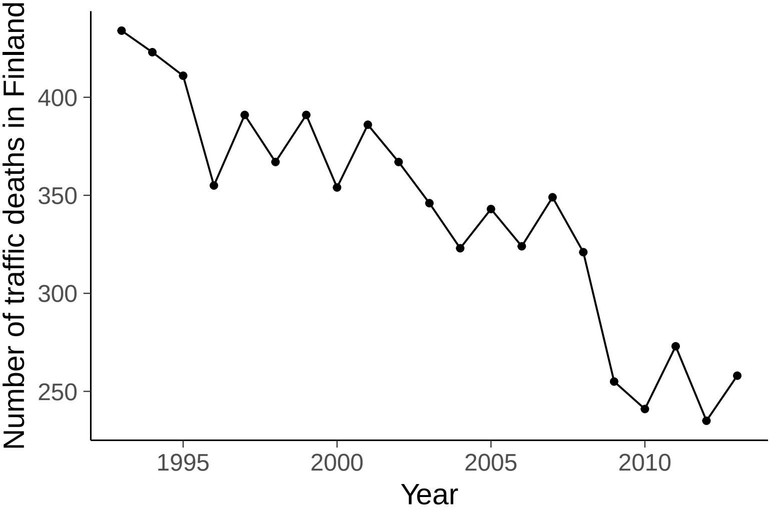
Poisson model for count data

- Number of traffic deaths per year (by Liikenneturva)



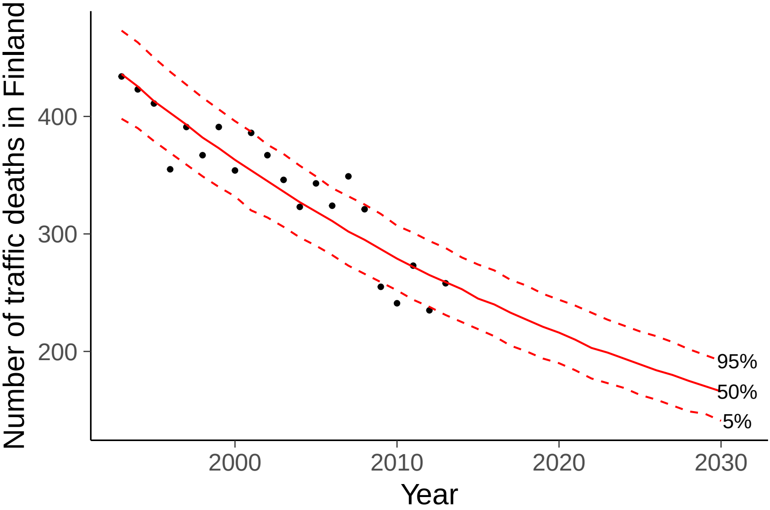
Poisson model for count data

- Number of traffic deaths per year (by Liikenneturva)



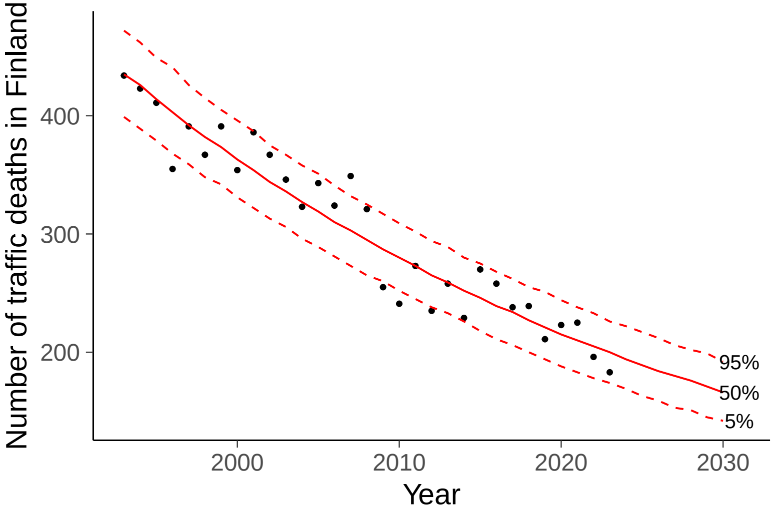
Poisson model for count data

- Number of traffic deaths per year (by Liikenneturva)



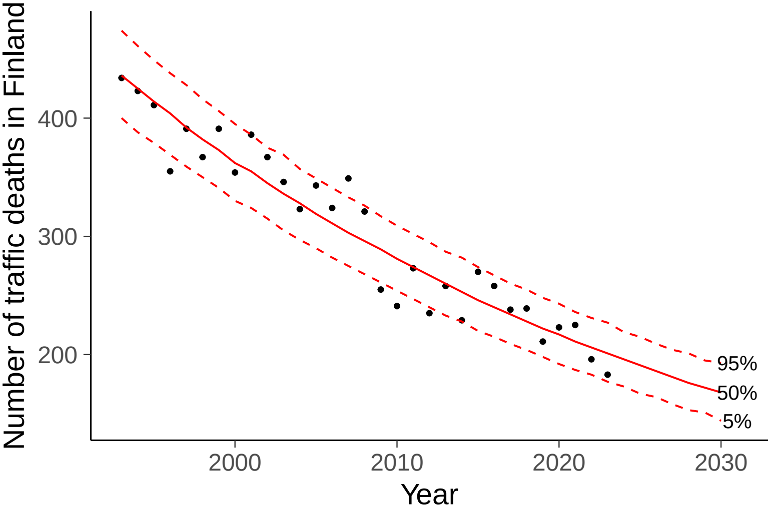
Poisson model for count data

- Number of traffic deaths per year (by Liikenneturva)



Poisson model for count data

- Number of traffic deaths per year (by Liikenneturva)



Thinking priors

- Make a guess of some quantities and then find out useful prior information for that. E.g.
  - proportion of students using MS Windows vs. Apple macOS vs. Linux
  - proportion of students who are taller than 1.9 m
  - proportion of students, who submitted the first assignment, attending the next lecture
  - proportion of students, who submitted the first assignment, submitting the last assignment