Fundamentals in Bayesian* Statistics

Malte Schierholz

Institute for Employment Research
University of Mannheim

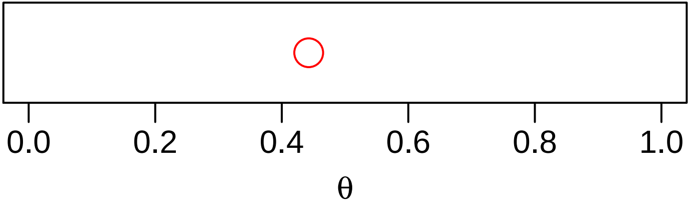
Dec 14, 2016

* Named after Reverend Thomas Bayes who published his famous paper about "inverse probabilities" posthumously in 1763

Example 1: Mathematical Bayes cont'd

Bayes (1763) considered the data generating process:

1. (Prior) Throw a ball at random on the table, $\theta \sim Unif(0,1)$
2. (Likelihood) Throw N further balls at random and count balls to the left, $N_{left}|\theta, N \sim Bin(N, \theta)$



If we knew N and N_{left} (but not θ), what can we say about θ?
Wanted:

$$Pr(\theta_1 < \theta < \theta_2 | N_{left}, N)$$

Example 1: Mathematical Bayes

Bayes' rule

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} \qquad \text{Def. conditional probability} \tag{1}$$

$$= \frac{p(y|\theta) \cdot p(\theta)}{p(y)} \qquad \text{Bayes' rule} \tag{2}$$

Application

$$Pr(\theta_1 < \theta < \theta_2 | N_{left}, N) = \int_{\theta_1}^{\theta_2} p(\theta | N_{left}, N) d\theta \tag{3}$$

$$= \int_{\theta_1}^{\theta_2} \frac{p(N_{left}|\theta, N) p(\theta)}{p(N_{left}|N)} d\theta \tag{4}$$

$$= \frac{\int_{\theta_1}^{\theta_2} \binom{N}{N_{left}} \theta^{N_{left}} (1-\theta)^{N-N_{left}} \cdot 1 d\theta}{\int_0^1 \binom{N}{N_{left}} \theta^{N_{left}} (1-\theta)^{N-N_{left}} d\theta} \tag{5}$$

Pure math (if we know the data generating process)! But it raises a century-long controversy how to interpret this result.

Example 2: Subjective Bayes

You are sitting in an airplane and everything is normal. Consider two hypotheses:

H_1: Flight is safe.

H_2: Flight has an emergency.

Your personal beliefs are:

$\mathbb{P}(H_1) = 0.9$ $\qquad$ $\mathbb{P}(H_2) = 1 - \mathbb{P}(H_1) = 0.1$

Example 2: Subjective Bayes

You are sitting in an airplane and everything is normal. Consider two hypotheses:
H_1: Flight is safe.
H_2: Flight has an emergency.

Your personal beliefs are:
$\mathbb{P}(H_1) = 0.9$ $\mathbb{P}(H_2) = 1 - \mathbb{P}(H_1) = 0.1$

Now, the pilot requests to fasten the seat belts (=observed data). You belief
$\mathbb{P}(Data|H_1) = 0.8$ (Minor turbulences)
$\mathbb{P}(Data|H_2) = 0.9$ (Sign of emergency)

What is the updated probability of an emergency?

Bayes Theorem (in its simplest form):

$$\mathbb{P}(H_i|Data) = \frac{\mathbb{P}(Data|H_i) \cdot \mathbb{P}(H_i)}{\sum_j \mathbb{P}(Data|H_j) \cdot \mathbb{P}(H_j)}$$

Example 2: Subjective Bayes cont'd

Your updated beliefs are:
$\mathbb{P}(H_1|belts) = 8/9 = 0.88$ (Flight is safe)
$\mathbb{P}(H_2|belts) = 1/9 = 0.11$ (Flight has an emergency)

Bayesian Updating
Suddenly, the oxygen masks drop (=observed data). You belief
$\mathbb{P}(masks|H_1) = 0.01$ (Technical error of mask system)
$\mathbb{P}(masks|H_2) = 0.9$ (Sign of emergency)

What is the updated probability of an emergency? Our posterior beliefs from above are prior beliefs for the current update:

$$\mathbb{P}(H_2|belts, masks) = \frac{\mathbb{P}(masks|H_2, belts) \cdot \mathbb{P}(H_2|belts)}{\sum_{j=1}^{2} \mathbb{P}(masks|H_j, belts) \cdot \mathbb{P}(H_j|belts)} \quad (6)$$

$$= \frac{0.9 \cdot 0.11}{0.01 \cdot 0.88 + 0.9 \cdot 0.11} = 91.8\% \quad (7)$$

Example 2: Subjective Bayes cont'd

The example shows some central features of Bayesianism:

- Probabilities are interpreted as personal degrees of belief.
- New data leads to belief updating
- Bayes' rule provides a mechanism to update personal beliefs *according to the laws of probability*
- Bayesian confirmation theory claims that one can update a Hypothesis when new evidence comes in:

$$\mathbb{P}(Hypothesis|Evidence) = \frac{\mathbb{P}(Evidence|Hypothesis) \cdot \mathbb{P}(Hypothesis)}{\mathbb{P}(Evidence)}$$

But even Bayesians do not agree with this view! (Gelman & Shalizi 2013)

Example 3: Proportion of Female Births

Laplace (1785) analyzed the proportion of female births in Europe.

- Data: 493,472 births (49% female) in Paris between 1745-1770

θ is the proportion of female births, a *random parameter*!

His Model:

- $N_{female}|\theta, N_{gesamt} \sim Bin(N_{gesamt}, \theta)$ (Likelihood)
- $\theta \sim Unif(0, 1)$ (Prior, all values are equally likely a priori)

His result:

$$Pr(\theta \geq 0.5 | N_{female} = 241,945, N_{gesamt} = 493,472) \approx 1.15 \times 10^{-42}$$

making him 'morally certain' that $\theta < 0.5$

(cited after Gelman et al. 2014, p. 31)

Controversy: Frequentist versus Bayesian Statistics

- Century-old controversy whose thinking is superior

Fundamental question: Are parameters θ fixed or random?

Frequentists:

- Parameters have a true value (although it is not directly observable). Probabilities about parameters are meaningless in this context (e.g. $\mathbb{P}(\theta = 1 | y_1 = 0, y_2 = 0)$).
  - Leads to frequentist interpretation of probabilities: Probabilities express relative frequencies when the random process were iterated ∞ times.

Bayesians:

- *All* unknown quantities, data and parameters, are random variables before they are observed.
  - Leads to subjective interpretation of probabilities: "I believe that ..."
- "The act of observation changes the status of the quantity from a random variable to a number." (Lindley 1975)
  - Parameters remain always uncertain. Probabilities express subjective degrees of belief.

A Different Paradigm

Forget everything you learned in your statistics classes:

- Classical estimation procedures (e.g., Maximum likelihood, OLS, Method of Moments)
- Properties of estimators, e.g.,
  - Unbiasedness: $\mathbb{E}_{\theta_0}(\hat{\theta}) = \theta_0$?
  - Consistency: For $n \to \infty$, is $\mathbb{P}(|\hat{\theta}_n - \theta_0| < \epsilon) = 1$?
  - MSE, Efficiency, ...
- Confidence intervals
  - Cover the true value θ_0 with frequency $(1 - \alpha)\%$
- Statistical testing and p-values
  - Answer the question: "if this hypothesis is true (which it might not be), what is the probability of observing even more extreme data (which we didn't)?"

Bayesians feel that they answer a more relevant question: "Given the observed data, what is the probability this hypothesis is true?"

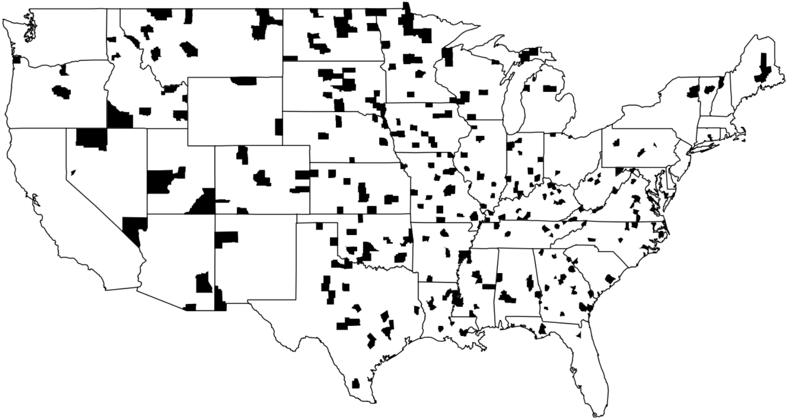
(Provocative quotes from Wolpert 2004)

Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

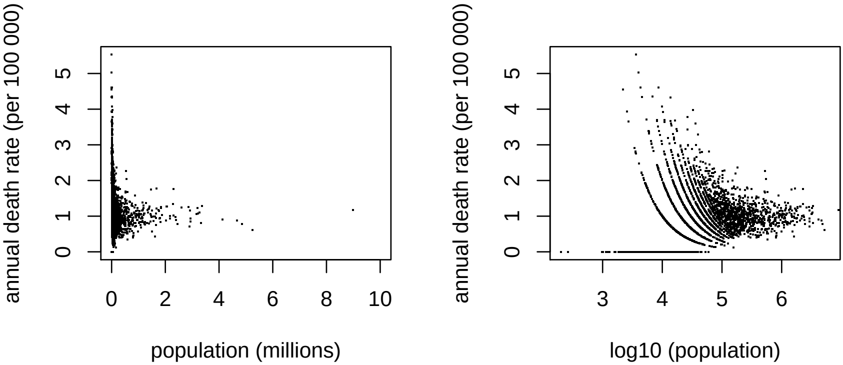
(see Gelman et al. 2014 for everything that follows)

Example: Kidney Cancer Death Rates



U.S. counties with 10% highest death rates for kidney cancer for U.S. males, 1980-1989
(graphic taken from Gelman et al. (2014))
Why are the highest death rates at the center of the map?

Example: Kidney Cancer Death Rates cont'd



Kidney cancer death rates $(y_j/10n_j)$ vs. population size n_j

Bayesian methods can help to calculate more accurate death rates, especially for small counties!

Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

Example cont'd: Model formulation

Idea: Estimate separate models for each county.

Notation:

- y_j: observed number of deaths in county j over 10 years
- θ_j: annual death rate in county j (per 100 000)
- n_j: population size in county j (in 100 000)
- $10 n_j \theta_j$: expected number of deaths in county j over 10 years

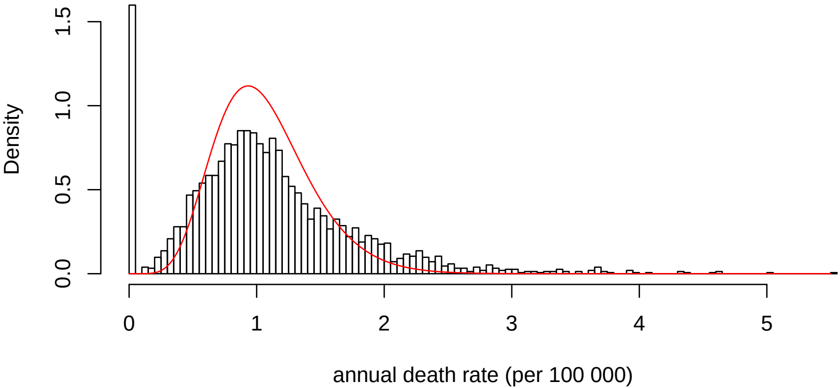
Model assumptions:

- Observed Likelihood:
  $y_j | \theta_j \sim \text{Poisson}(10 n_j \theta_j)$
  $\Rightarrow y_j | \theta_j$ has density $p(y_j | \theta_j) = \frac{(10 n_j)^{y_j}}{y_j!} \theta_j^{y_j} \exp(-10 n_j \theta_j)$
- Prior:
  $\theta_j \sim \text{Gamma}(\alpha = 8, \beta = 7.5)$
  $\Rightarrow \theta_j$ has density $p(\theta_j) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_j^{\alpha - 1} \exp(-\beta \theta_j)$

Example cont'd: Prior Distribution

Where does this prior come from?

- Gamma distribution was chosen for computational convenience
- With parameters $\alpha = 8, \beta = 7.5$ the distribution is similar to the observed death rates



Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

Calculation and Interpretation

1. Calculations: Central goal is to get access to the posterior distribution $p(\theta|y)$:

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

An equivalent form omits all factors that do not depend on θ, yielding the *unnormalized posterior density*

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \textbf{Likelihood} \times \textbf{prior} \end{aligned}$$

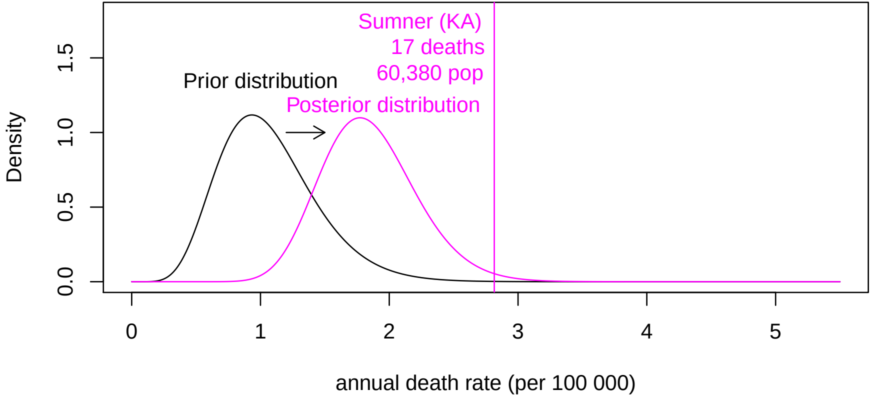
which simplifies subsequent calculations.

2. Interpretation: What does this posterior density tell us?

Example cont'd: Posterior distribution

- Prior:
  $\theta_j \sim \text{Gamma}(\alpha = 8, \beta = 7.5)$
- Posterior distribution for $\alpha = 8$, $\beta = 7.5$:
  $\theta_j | y_j \sim \text{Gamma}(8 + y_j, 7.5 + 10n_j)$

What does this posterior look like?



Example cont'd: Model Summaries and Interpretation

Several summaries of the distribution are helpful for interpretation:

- Point Estimates:
  - **Posterior mean** $= \mathbb{E}(\theta_j | y_j) = 1.85$
  - **Posterior median** $= \text{Median}(\theta_j | y_j) = 1.82$
  - **Posterior mode** $= \text{Mode}(\theta_j | y_j) = 1.77$
- The 95% central **posterior interval** covers the interval between the 0.025-quantile and the 0.975-quantile. Here: $\mathbb{P}(1.19 < \theta_j < 2.63) = 0.95$
- ...

The reasoning to calculate the numbers above is:

1. Assume a Gamma prior to obtain a Gamma posterior
2. Closed formulas to calculate the posterior mean and the quantiles exists only for some distributions (like Gamma).

If a different prior were assumed, no closed formulas exist to calculate the summaries. More complex computational techniques (MCMC) are needed instead.

Interpreting the prior

The choice of priors is important! But what is a substantial interpretation for the prior $\theta_j \sim \text{Gamma}(\alpha = 8, \beta = 7.5)$ we choose in our example?

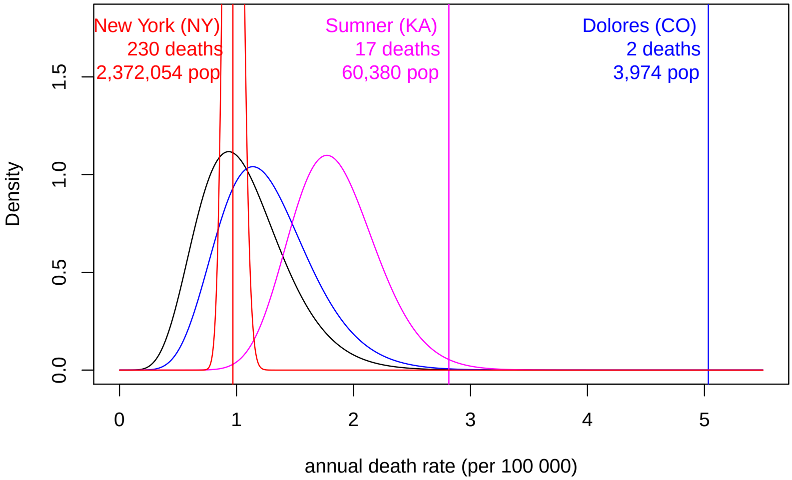
Sumner (KA) has posterior distribution

$$\theta_j | y_j \sim Gamma(8 + 17, 7.5 + 10 \cdot 0.6038)$$

It remains the same if additional data were observed but prior parameters were decreased by the same amount, suggesting the interpretation:

- From an imaginary prior study we know that $\alpha = 8$ persons out of a population of size $\beta = 7.5$ (in 100 000s) died.

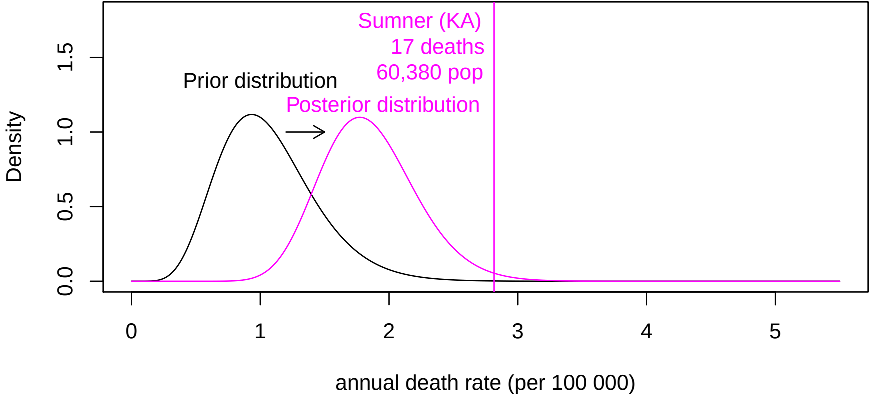
Example cont'd: Belief Updating



Again: Posterior distributions are in between the prior distribution and the observed data. Large counties are more influential.

Example cont'd: Predictions

Posterior distribution for $\alpha = 8$, $\beta = 7.5$:
$\theta_j | y_j \sim \text{Gamma}(8 + y_j, 7.5 + 10 n_j)$



How many people will die in Sumner next year?

- Frequentist plug-in solution: $n_{Sumner} \cdot \hat{\theta}_{Sumner}$, e.g. $n_{Sumner} \cdot \mathbb{E}(\theta_j | y_j) = 0.6038 \cdot 1.85 = 1.12$ deaths per year
- Uncertainty about θ_j is ignored

Predictive Distributions

Predictions can be based on the prior distribution or on incoming data:

Prior predictive distribution for an observation y (not conditional on observed data):

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

- Has applications in model comparison and model averaging

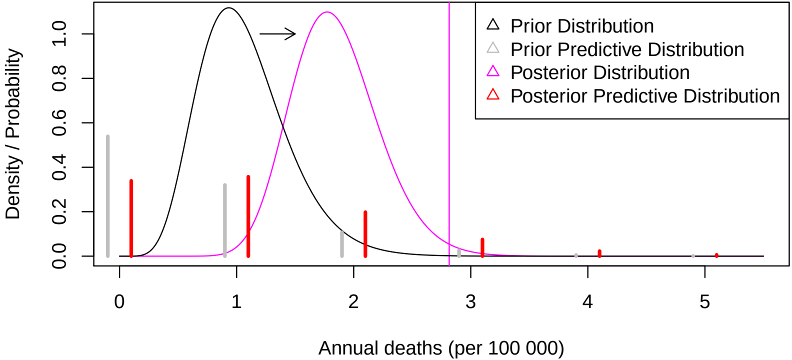
Posterior predictive distribution for a future observation y_f (conditional on observations y):

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

Predictive distributions are weighted means of conditional predictions $p(y_f|\theta)$ with weights $p(\theta|y)$

Example cont'd: Predictive Distributions

How would a Bayesian predict how many people die the next year in Sumner?



After observing the data, $\mathbb{P}(y_f = 0|y) = 0.35$ which is lower than our prior predicted probability.

Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

Evaluate the Model Fit

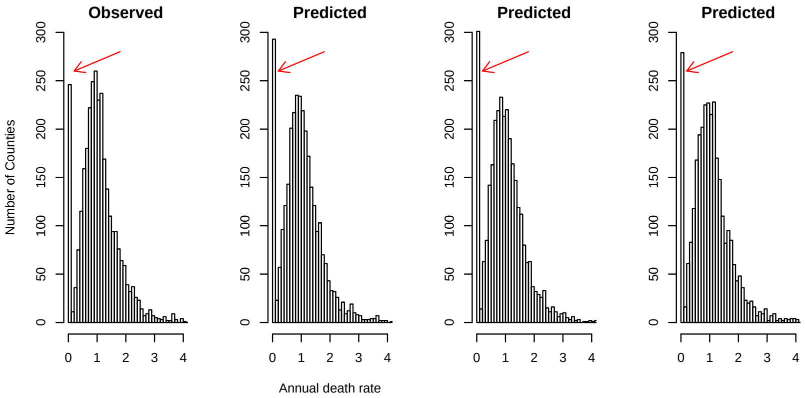
Procedure:

- Make a single prediction from the posterior predictive distribution for all observations (all counties).
- If the model is adequate to generate the data, the distributions of observed and predicted data should be similar
- Generate multiple predictions to see what typical predictions look like
- *Bayesian p-values* are defined as the probability that the replicated data is more extreme than the observed data,

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

Example cont'd: Evaluate the Model Fit

How well does our model replicate the observed data?



Our model typically generates more counties with death rate $= 0$ than what has been observed.

In fact, it always does so, $p_B = 100\%$! Revise model.

Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

A note on notation:
So far we considered only models with a single parameter, the death rate $\theta = \theta_i$. From now on we look at models with many parameters and θ is meant to subsume all the unknown parameters, $\theta = (\theta_1, \theta_2, ..., \theta_J, \alpha, \beta)$

Example cont'd: Prior Distribution Revisited

Prior in the kidney example was somehow ad-hoc: There is no good reason why I chose parameters $\alpha = 8$, $\beta = 7.5$.

What are other strategies to estimate the death rate in Sumner?

1. Only use data from this county:
   Death rate $= \frac{y_j}{10 n_j} = 2.82$
2. Calculate death rate for the U.S.
   Death rate $= \frac{\sum y_j}{10 \sum n_j} = 1.01$
   and assume that this rate is identical in all counties

The posterior mean $\mathbb{E}(\theta_j | y_j) = 1.85$ from the model above is a compromise between both extremes. Other priors would yield other estimates. For example:

- A prior with infinite variance leads to strategy 1 (e.g. $Gamma(0, 0)$)
- A prior with zero variance centered at 1.01 leads to strategy 2 (e.g. $Gamma(\alpha \to \infty, \beta \to \infty)$ with $\frac{\alpha}{\beta} = 1.01$)

Hierarchical Models

Can we find a better compromise between the local (county-level) and the global (country-level) model? That is, how can we use data from other counties to find a better prior for Sumner?

Hierarchical models do exactly this. Central assumptions:

1. County-level death rates $\theta_i, i = 1, ..., J$, are all drawn from a common *population distribution*, $\theta_i|\alpha, \beta \sim Gamma(\alpha, \beta)$.
   - Same distribution as the prior before but now inference happens in parallel for all counties.
   - Why are all θ_i from the same distribution? Because they are *exchangeable*: Before looking at the data, nothing is known about differences between death rates.
2. The population distribution $Gamma(\alpha, \beta)$ will be estimated from the data.
   - Parameters α, β are therefore uncertain (= random variables). They need a *hyperprior distribution* $p(\alpha, \beta)$

Formally, prior and posterior distributions take a new form:

- Prior: $p(\theta_1, ..., \theta_J, \alpha, \beta) = p(\theta_1, ..., \theta_J|\alpha, \beta)p(\alpha, \beta)$
- Posterior: $p(\theta_1, ..., \theta_J, \alpha, \beta|y) \propto p(y|\theta, \alpha, \beta)p(\theta, \alpha, \beta)$

Example cont'd: Hierarchical Models

How can we specify the prior?
$p(\theta_1, ..., \theta_J, \alpha, \beta) = p(\theta_1, ..., \theta_J | \alpha, \beta) p(\alpha, \beta)$

Assumptions:

- $p(\theta_1, ..., \theta_J | \alpha, \beta) = \prod p(\theta_i | \alpha, \beta)$ with $\theta_i | \alpha, \beta \sim Gamma(\alpha, \beta)$
- But which density for $p(\alpha, \beta)$?
  - Reparametrize the gamma distribution in terms of country-level mean μ and standard deviation σ: $\alpha = (\mu/\sigma)^2$ and $\beta = \mu/\sigma^2$
  - $p_{\alpha,\beta}(\alpha, \beta) = p_{\mu,\sigma}(\mu, \sigma) = p_\mu(\mu) p_\sigma(\sigma)$, assuming $\mu \perp \sigma$
  - $f(\mu) \propto constant$, because all possible death rates should be equally likely
  - $\sigma \sim Uniform(0, 10)$, because the variance must be > 0 and data shows that it is small
- This prior is uninformative and robust. Results are not sensitive to the hyperprior distribution $p_{\mu,\sigma}(\mu, \sigma)$.

Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

Example cont'd: Calculations

Analytic derivations of posterior quanitities of interest is hard or impossible. *Markov Chain Monte Carlo* simulations (MCMC) are often used to gain insights about the posterior.

MCMC simulation is here implemented with the software Stan:

```
  real<lower=0> theta[J];
}
transformed parameters {
  real<lower=0> alpha;
  real<lower=0> beta;
  real<lower=0> expdeaths[J]; // expected number of deaths in county j

  alpha <- (expectation / sde)^2;
  beta <- expectation / (sde^2);

  for (j in 1:J)
    expdeaths[j] <- n[j] * theta[j] * 10; # in ten years
 }
model {
  sde ~ uniform(0, 10);

  theta ~ gamma(alpha, beta);
  y ~ poisson(expdeaths);
}
```

Markov Chain Monte Carlo simulations

Basic idea for simulation-based techniques:

1. Draw S random numbers $\theta^{(1)}, ..., \theta^{S}$ from the posterior $p(\theta|y)$

   MCMC algorithms do this as follows:

   - Start with an arbitrary starting point $\theta^{(0)}$
   - For $s = 1, 2, ...$:
     - Sample a proposal θ^* from a well-suited *proposal distribution* $J_s(\theta^*|\theta^{(s-1)})$ that must depend on the latest draw $\theta^{(s-1)}$
     - Set

       $$\theta^{(s)} = \begin{cases} \theta^* \text{ with probability } p \text{ that is calculated from } p(\theta|y) \text{ and } J \\ \theta^{(s-1)} \text{ otherwise} \end{cases}$$

   Theorem: The sequence of iterations $\theta^{(1)}, \theta^{(2)}, ...$ converges to the posterior $p(\theta|y)$

2. Calculate the quantities of interest from $\theta^{(1)}, ..., \theta^{(S)}$, e.g.:
   - Posterior mean $= \mathbb{E}(\theta|y) \approx \frac{1}{S} \sum_{s=1}^{S} \theta^{(s)}$
   - Posterior median $=$ Median$(\theta|y) \approx$ After sorting $\theta^{(1)}, ..., \theta^{(S)}$, pick the middle one.

Markov Chain Monte Carlo simulations

Basic idea for simulation-based techniques:

1. Draw S random numbers $\theta^{(1)}, ..., \theta^{S}$ from the posterior $p(\theta|y)$

   MCMC algorithms do this as follows:

   - Start with an arbitrary starting point $\theta^{(0)}$
   - Iterate for $s = 1, 2, ...$:
     - Sample a proposal θ^* from a well-suited *proposal distribution* $J_s(\theta^*|\theta^{(s-1)})$ that must depend on the latest draw $\theta^{(s-1)}$
     - Set

$$\theta^{(s)} = \begin{cases} \theta^* \text{ with probability } p \text{ that is calculated from } p(\theta|y) \text{ and } J \\ \theta^{(s-1)} \text{ otherwise} \end{cases}$$

   Theorem: The sequence of iterations $\theta^{(1)}, \theta^{(2)}, ...$ converges to the posterior $p(\theta|y)$

2. Calculate the quantities of interest from $\theta^{(1)}, ..., \theta^{(S)}$, e.g.:
   - Posterior mean $= \mathbb{E}(\theta|y) \approx \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}$
   - Posterior median $=$ Median$(\theta|y) \approx$ After sorting $\theta^{(1)}, ..., \theta^{(S)}$, pick the middle one.

MCMC Diagnostics

Two challenges arise when using MCMC.
Diagnostics needed:

- The first values $\theta^{(1)}, \theta^{(2)}, \ldots$ depend on the starting value $\theta^{(0)}$ and are not representative for the posterior $p(\theta|y)$.
  Remedies:
  - Discard the first half of the sequence $\theta^{(1)}, \ldots, \theta^{(S)}$
  - Check that the second half has reached convergence
- Subsequent simulation draws $\theta^{(s-1)}$ and $\theta^{(s)}$ are correlated.
  This reduces the *effective* number of draws.
  Remedy:
  - Calculate the effective sample size and make sure it is large enough for follow-up calculations.

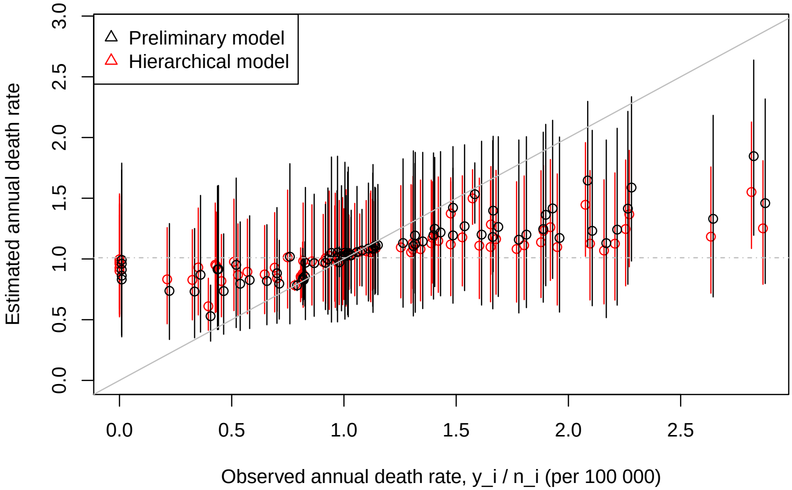
Example cont'd: Diagnostics

Example cont'd: Results

Parameter	mean	2.5%	50%	97.5%
μ	1.035	1.018	1.035	1.053
σ	0.262	0.244	0.262	0.280
α	15.665	13.831	15.649	17.842
β	15.133	13.318	15.105	17.339
θ_{Sumner}	1.551	1.086	1.534	2.127
$\theta_{Dolores}$	1.140	0.671	1.129	1.712
$\theta_{NewYork}$	0.974	0.858	0.973	1.105

Parameters $\alpha = 8, \beta = 7.5$ in the preliminary analysis above were too low!
$\mathbb{P}(\theta_{Dolores} > \theta_{Sumner}) = 13.4\%$ compared to 14.2% in the preliminary model (difference might be due to Monte Carlo simulations).

Example cont'd: Results for 80 Random Counties



- Both models shrink the estimates towards the mean (stronger for the hierarchical model)
- Hierarchical model has shorter 0.95-credibility intervals

Three Steps in Bayesian Data Analysis

1. Set up the probability model. This is a joint probability distribution over all observable and unobservable variables. It includes
   - the *likelihood function* $p(y|\theta)$ that describes how the data was generated given the unknown quantity θ.
   - the *prior* density $p(\theta)$ that describes prior knowledge about θ.
2. Calculate and interpret the posterior density $p(\theta|y)$.
3. Evaluate the model fit: How well does the model fit the data and how sensitive are the results to modeling assumptions in step 1?
   - Repeat the three steps if the assumptions from step 1 are not satisfactory.

Example cont'd: Evaluate the Model Fit

How well would this model predict the observed data?

- The hierarchical model still predicts too many counties with death rate $= 0$.
- There are also too many counties with high predicted death rates. $\Rightarrow$ Overdispersion
- Though, the hierarchical model is better than the original model according to both criteria.

One might try to improve the model further:

- Information at the county-level might explain differences between counties. Technique for analysis: Hierarchical poisson regression
- Spatial models

Interim Summary

Bayesian Statistics ...

- ... is different to Frequentist Statistics by treating all unknown quantities as random
- ... is all about the posterior distribution and its implications,
- ... emphasizes uncertainty in the form of distributions and credibility intervals,
- ... emphasizes the need for model checking and sensitivity analysis,
- ... allows for models with many parameters,
- ... requires explicit assumptions about the data generation and about the prior distribution,
- ... is now easier than ever before because MCMC-sampling can usually replace cumbersome analytic calculations for the posterior distribution.

Let's play soccer! - What can Bayesian Statistics do?

Bååth (2013) modeled the number of goals for the Spanish La Liga. See his blog post for details and additions.

Model assumptions:

$$\begin{aligned} HomeGoals_{ij} &\sim Poisson(\lambda_{home,ij}) && (8)\\ AwayGoals_{ij} &\sim Poisson(\lambda_{away,ij}) && (9)\\ \log(\lambda_{home,ij}) &= baseline_{home} + skill_i + skill_j && (10)\\ \log(\lambda_{home,ij}) &= baseline_{away} + skill_i + skill_j && (11) \end{aligned}$$

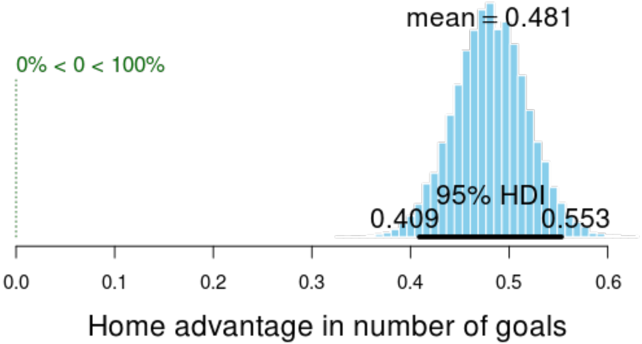
Prior assumptions:

$$\begin{aligned} baseline &\sim Normal(0, 4^2) && (12)\\ skill_{1\ldots n} &\sim Normal(\mu_{teams}, \sigma^2_{teams}) && (13)\\ \mu_{teams} &\sim Normal(0, 4^2) && (14)\\ \sigma_{teams} &\sim Uniform(0, 3) && (15) \end{aligned}$$

Add-on: Analyze more than one season:

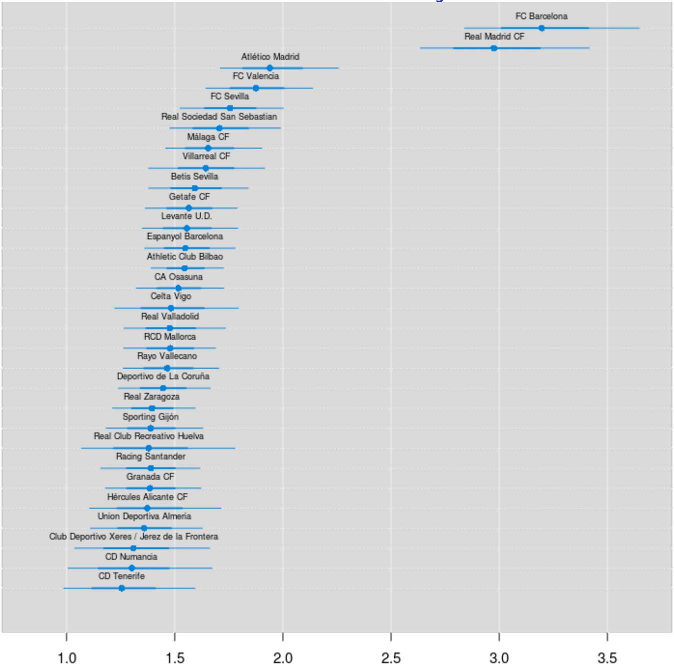
$$skill_{i,t+1} \sim Normal(skill_{i,t}, \sigma^2_{season}) \qquad (16)$$

Soccer Results I



$$p(\exp(baseline_{home}) - \exp(baseline_{away}) | Goals)$$

Soccer Results II: Rank teams by skill level



Soccer Results III: Predict number of goals for various matches

HomeTeam	AwayTeam	mean_home_goals	mean_away_goals	mode_home_goals	mode_away_goals
Celta Vigo	Athletic Club Bilbao	1.50	1.10	1.00	1.00
Deportivo de La Coruña	Atlético Madrid	1.20	1.50	1.00	1.00
FC Barcelona	Betis Sevilla	3.20	0.50	3.00	0.00
FC Sevilla	Espanyol Barcelona	1.80	1.00	1.00	0.00
FC Valencia	CA Osasuna	2.00	0.90	1.00	0.00
Getafe CF	Real Sociedad San Sebastian	1.40	1.20	1.00	1.00
Granada CF	Málaga CF	1.30	1.30	1.00	1.00
RCD Mallorca	Levante U.D.	1.50	1.10	1.00	1.00
Real Madrid CF	Real Valladolid	3.20	0.50	3.00	0.00
Real Zaragoza	Rayo Vallecano	1.50	1.10	1.00	1.00
Athletic Club Bilbao	RCD Mallorca	1.60	1.00	1.00	1.00
Atlético Madrid	FC Barcelona	1.00	1.80	0.00	1.00
Betis Sevilla	Celta Vigo	1.70	1.00	1.00	1.00
CA Osasuna	Getafe CF	1.50	1.10	1.00	1.00
Espanyol Barcelona	Real Madrid CF	0.80	2.10	0.00	1.00
Levante U.D.	Real Zaragoza	1.80	1.00	1.00	0.00
Málaga CF	FC Sevilla	1.50	1.10	1.00	1.00
Rayo Vallecano	FC Valencia	1.20	1.40	1.00	1.00
Real Sociedad San Sebastian	Granada CF	2.00	0.90	1.00	0.00
Real Valladolid	Deportivo de La Coruña	1.60	1.10	1.00	1.00
Celta Vigo	Atlético Madrid	1.20	1.40	1.00	1.00
Deportivo de La Coruña	Espanyol Barcelona	1.50	1.20	1.00	1.00
FC Barcelona	Real Valladolid	3.40	0.50	3.00	0.00
FC Sevilla	Real Sociedad San Sebastian	1.60	1.10	1.00	1.00

Some Notes about Priors

Critics of Bayesian methods often question prior assumptions. After all, we want to infer knowledge from the data - not reconfirm the prior illusions.

- For testing a hypothesis, the prior should certainly not favor this hypothesis.

There are various philosophies on prior specification:

- **Informative priors** are constructed from expert knowledge or with estimates from the literature.
- **Noninformative priors** are those that have minimal influence on the posterior, e.g. the flat prior $p(\theta) \propto cons$ on $(-\infty, \infty)$
- **Weakly informative priors** use a tiny bit of actual prior knowledge, just enough to avoid mathematical difficulties with noninformative priors.
  - Example: A prior for the sex ratio at birth could be concentrated between 0.4 and 0.6

Relation to Maximum-Likelihood Estimation

There exists a close correspondence between Bayesian Statistics and Maximum-Likelihood Estimation:

- $\hat{\theta}_{ML} = \text{Modus}(\theta|y)$ if the prior is noninformative, $p(\theta) \propto cons$

Under all reasonable prior distributions, results become also more similar to $\hat{\theta}_{ML}$ when more data arrives:

- With increasing sample size, $n \to \infty$, the posterior mode $\hat{\theta}_{mode}$ and the ML-estimate $\hat{\theta}_{ML}$ are both *consistent*, *asymptotically unbiased* and converge to the same *normal distribution*:

  $$\mathbb{P}(\theta|y) \approx \mathbb{P}(\hat{\theta}_{ML}(y)) \approx Normal(\theta_0, I^{-1}(\hat{\theta}))$$

  with "true" parameter θ_0 and inverse information matrix $I^{-1} \stackrel{n\to\infty}{\longrightarrow} 0$.

Summary & Outlook

For many scenarios there exist Bayesian and Frequentist solutions that often - but not always - lead to the same conclusion.

Bayesian Statistics can be advantageous because ...

- ... results improve on small data sets when meaningful prior information is available, and converge to frequentist solutions for large data sets,
- ... it provides techniques to combine information from different sources and complicated data structures,
- ... it can motivate solutions for various problems (e.g., multiple imputation),
- ... it simplifies interpretation of results (credibility vs. confidence intervals).

But Bayesian Statistics also ...

- ... requires explicit assumptions about a prior distribution,
- ... requires skills how to carry out MCMC-simulations.

References

Bååth, R. (2013). *Modeling Match Results in La Liga Using a Hierarchical Bayesian Poisson Model*. Blog post. Online at `http://www.sumsar.net/blog/2013/07/modeling-match-results-in-la-liga-part-one/`

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, pp. 330-418.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, third edition. New York: Chapman & Hall.

Gelman A. and Shalizi, C. (2013). Philosophy and the practice of Bayesian statistics (with discussion). *British Journal of Mathematical and Statistical Psychology* **66**, 8-80

References

Lindley, D.V. (1975). The Future of Statistics: A Bayesian 21st century. *Supp. Advances in Applied Probability* **7**, pp. 106-115.

Wolpert, R. L. (2004). A Conversation with James O. Berger. *Statistical Science* **19**, Special issue on Bayes Then and Now, pp. 205-218.

Example: Hemophilia

- *Background*: Hemophilia is a blood disease that is inherited via the X-chromosome. Males are affected when they inherit a single chromosome with genetic disorder, women are only affected when both X-chromosomes have this disorder.
- *Prior Information*: A woman's brother was affected by the disease implying that her mother was a carrier of the disease. Her father did not have the disease. What is the probability that the woman herself is also a carrier of the disease ($\theta = 1$) or not ($\theta = 0$)? With this information it is equally likely that she has inherited the chromosome in disorder or not, $\mathbb{P}(\theta = 1) = \mathbb{P}(\theta = 0) = 0.5$.
- *Observed data and model*: Suppose the woman has two sons, neither of whom is affected ($y_1 = y_2 = 0$). Conditional on θ, the outcomes of both are independent. The likelihood is:

$$\mathbb{P}(y_1 = 0, y_2 = 0|\theta = 0) = \mathbb{P}(y_1 = 0|\theta = 0)\,\mathbb{P}(y_2 = 0|\theta = 0) = 1 * 1$$
$$\mathbb{P}(y_1 = 0, y_2 = 0|\theta = 1) = \mathbb{P}(y_1 = 0|\theta = 1)\,\mathbb{P}(y_2 = 0|\theta = 1) = ?$$

Find the updated posterior probability that the woman is a carrier of the disease, $\mathbb{P}(\theta = 1|y_1 = 0, y_2 = 0)$.

Example: Hemophilia cont'd

$$\mathbb{P}(\theta = 1) = \mathbb{P}(\theta = 0) = 0.5$$
$$\mathbb{P}(y_1 = 0, y_2 = 0|\theta = 0) = 1$$
$$\mathbb{P}(y_1 = 0, y_2 = 0|\theta = 1) = 0.5 * 0.5 = 0.25$$

- Remember Bayes Theorem:

$$\mathbb{P}(\theta|y_1, y_2) = \frac{\mathbb{P}(\theta, y_1, y_2)}{\mathbb{P}(y_1, y_2)} = \frac{\mathbb{P}(y_1, y_2|\theta)\,\mathbb{P}(\theta)}{\mathbb{P}(y_1, y_2)}$$

Applying this formula one can easily calculate the posterior probability:

$$\mathbb{P}(\theta = 1|y) = \frac{\mathbb{P}(y_1 = 0, y_2 = 0|\theta = 1)\,\mathbb{P}(\theta = 1)}{\mathbb{P}(y|\theta = 1)\,\mathbb{P}(\theta = 1) + \mathbb{P}(y|\theta = 0)\,\mathbb{P}(\theta = 0)}$$
$$= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1)(0.5)} = \frac{0.125}{0.625} = 0.20$$

Example: Hemophilia cont'd

Suppose the woman has a third son who is also not affected ($y_3 = 0$). How can we update our belief if the woman is a carrier? Calculations from above:

$$\begin{aligned}
\mathbb{P}(\theta = 1 | y_1 = 0, y_2 = 0) &= 0.2 \\
\mathbb{P}(\theta = 0 | y_1 = 0, y_2 = 0) &= 0.8 \\
\mathbb{P}(y_3 = 0 | \theta = 1) &= 0.5 \\
\mathbb{P}(y_3 = 0 | \theta = 0) &= 1
\end{aligned}$$

Bayesian Updating: Adding more data to the analysis is straightforward. We don't need to do the calculations again. Instead, we simply use the posterior distribution from above as our new prior distribution.

$$\begin{aligned}
\mathbb{P}(\theta = 1 | y_1, y_2, y_3) &= \frac{\mathbb{P}(y_3 = 0 | \theta = 1, (y_1, y_2))\, \mathbb{P}(\theta = 1 | y_1, y_2)}{\mathbb{P}(y_3 | y_1, y_2)} \\
&= \frac{(0.5)(0.2)}{(0.5)(0.2) + (1)(0.8)} = 0.111
\end{aligned}$$

Combining Data and Prior Information

Bayesian updating is a central feature of Bayesian Statistics:

- The prior belief (based on the woman's family background)
- is updated based on data (her first two sons),
- and additional data allows further updating (her third son).

The resulting posterior belief is always a compromise between the prior belief and the observed data.

What would your conclusion be if you had only known the

- family background, or,
- alternatively, data from non-affected sons?

Bayesian Statistics allows combining different sources of knowledge!