Bayesian Inference with Sets of Conjugate Priors

Gero Walter

Department of Statistics
Ludwig-Maximilians-Universität München (LMU)

June 29th, 2012

Introduction

- Bernoulli observations: 0/1 observations (team wins no/yes)

Introduction

- Bernoulli observations: 0/1 observations (team wins no/yes)
- given: a set of observations (team won 12 out of 16 matches)

Introduction

- Bernoulli observations: 0/1 observations (team wins no/yes)
- given: a set of observations (team won 12 out of 16 matches)
- additional to observations, we have strong prior information (we are convinced that $P(\text{win})$ should be around 0.75)

Introduction

- Bernoulli observations: 0/1 observations (team wins no/yes)
- given: a set of observations (team won 12 out of 16 matches)
- additional to observations, we have strong prior information (we are convinced that $P(\text{win})$ should be around 0.75)
- we are, e.g., interested in (predictive) probability P that team wins in the next match

Introduction

- Bernoulli observations: 0/1 observations (team wins no/yes)
- given: a set of observations (team won 12 out of 16 matches)
- additional to observations, we have strong prior information (we are convinced that $P(\text{win})$ should be around 0.75)
- we are, e.g., interested in (predictive) probability P that team wins in the next match
- standard statistical model for this situation: Beta-Bernoulli/Binomial Model

Beta-Bernoulli/Binomial Model (BBM)

- Beta prior on $p = P(\text{win})$
- here in parameterization used, e.g., by Walley (1991):

Beta-Bernoulli/Binomial Model (BBM)

- Beta prior on $p = P(\text{win})$
- here in parameterization used, e.g., by Walley (1991):

$$
\begin{array}{r|ccl}
\text{Data :} & s & \sim & \text{Binom}(p, n) \\
\text{conjugate prior:} & p & \sim & \text{Beta}(n^{(0)}, y^{(0)}) \\
\hline
\text{posterior:} & p \mid s & \sim & \text{Beta}(n^{(n)}, y^{(n)})
\end{array}
$$

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{s}{n}, \qquad n^{(n)} = n^{(0)} + n$$

Beta-Bernoulli/Binomial Model (BBM)

- Beta prior on $p = P(\text{win})$
- here in parameterization used, e.g., by Walley (1991):

$$\begin{array}{r|ccl} \text{Data}: & s & \sim & \text{Binom}(p, n) \\ \text{conjugate prior}: & p & \sim & \text{Beta}(n^{(0)}, y^{(0)}) \\ \hline \text{posterior}: & p \mid s & \sim & \text{Beta}(n^{(n)}, y^{(n)}) \end{array}$$

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{s}{n}, \qquad n^{(n)} = n^{(0)} + n$$

$$y^{(n)} = \mathrm{E}[p \mid s] \qquad \mathrm{Var}(p \mid s) = \frac{y^{(n)}(1 - y^{(n)})}{n^{(n)} + 1}$$

Beta-Bernoulli/Binomial Model (BBM)

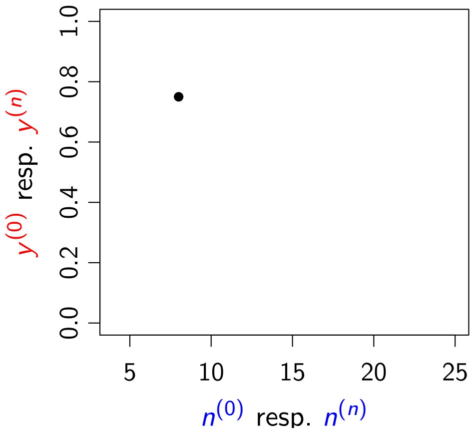
- Beta prior on $p = P(\text{win})$
- here in parameterization used, e.g., by Walley (1991):

Data :	s	$\sim$	$\text{Binom}(p, n)$
conjugate prior:	p	$\sim$	$\text{Beta}(n^{(0)}, y^{(0)})$
posterior:	$p \mid s$	$\sim$	$\text{Beta}(n^{(n)}, y^{(n)})$

$$y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{s}{n}, \qquad n^{(n)} = n^{(0)} + n$$

$$y^{(n)} = \text{E}[p \mid s] = \text{P} \qquad \text{Var}(p \mid s) = \frac{y^{(n)}(1 - y^{(n)})}{n^{(n)} + 1}$$

Beta-Bernoulli/Binomial Model (BBM)

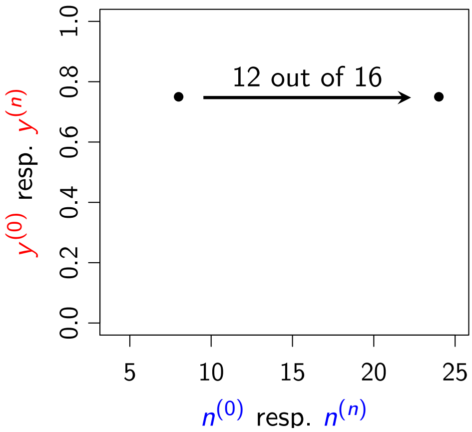


no conflict:

prior $n^{(0)} = 8$, $y^{(0)} = 0.75$

data $s/n = 12/16 = 0.75$

Beta-Bernoulli/Binomial Model (BBM)

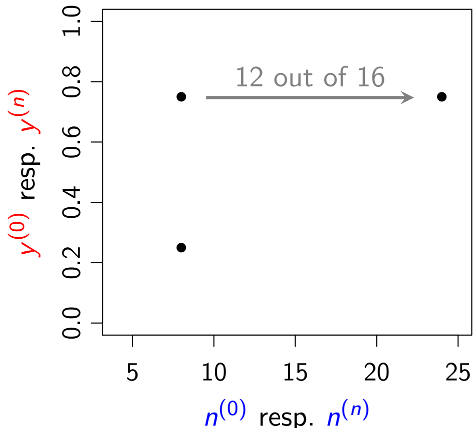


no conflict:

prior $n^{(0)} = 8$, $y^{(0)} = 0.75$
data $s/n = 12/16 = 0.75$

$n^{(n)} = 24$, $y^{(n)} = 0.75$

Beta-Bernoulli/Binomial Model (BBM)



no conflict:

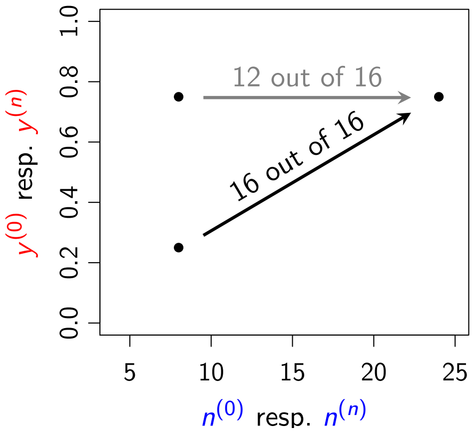
prior $n^{(0)} = 8$, $y^{(0)} = 0.75$
data $s/n = 12/16 = 0.75$

$n^{(n)} = 24$, $y^{(n)} = 0.75$

prior-data conflict:

prior $n^{(0)} = 8$, $y^{(0)} = 0.25$
data $s/n = 16/16 = 1$

Beta-Bernoulli/Binomial Model (BBM)



no conflict:

prior $n^{(0)} = 8$, $y^{(0)} = 0.75$
data $s/n = 12/16 = 0.75$

$n^{(n)} = 24$, $y^{(n)} = 0.75$

prior-data conflict:

prior $n^{(0)} = 8$, $y^{(0)} = 0.25$
data $s/n = 16/16 = 1$

Prior-Data Conflict $\hat{=}$ situation in which...

- ... *informative prior beliefs* and *trusted data* (sampling model correct, no outliers, etc.) are in conflict.
- "...the prior [places] its mass primarily on distributions in the sampling model for which the observed data is surprising." (Evans & Moshonov, 2006)
- ... there are not enough data to overrule the prior.

We should notice prior-data conflict in the posterior.

Prior-Data Conflict $\hat{=}$ situation in which...

- ...*informative prior beliefs* and *trusted data* (sampling model correct, no outliers, etc.) are in conflict.
- "...the prior [places] its mass primarily on distributions in the sampling model for which the observed data is surprising." (Evans & Moshonov, 2006)
- ...there are not enough data to overrule the prior.

We should notice prior-data conflict in the posterior.

$$\mathrm{E}[p \mid s] = y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{s}{n}$$

➡ Conflict between prior and data is just averaged out!

Prior-Data Conflict $\hat{=}$ situation in which...

- ...*informative prior beliefs* and *trusted data* (sampling model correct, no outliers, etc.) are in conflict.
- "...the prior [places] its mass primarily on distributions in the sampling model for which the observed data is surprising." (Evans & Moshonov, 2006)
- ...there are not enough data to overrule the prior.

We should notice prior-data conflict in the posterior.

$$\mathrm{E}[p \mid s] = y^{(n)} = \frac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \frac{n}{n^{(0)} + n} \cdot \frac{s}{n}$$

→ Conflict between prior and data is just averaged out!

$$\mathrm{Var}(p \mid s) = \frac{y^{(n)}(1 - y^{(n)})}{n^{(n)} + 1}, \qquad n^{(n)} = n^{(0)} + n$$

→ does not change systematically with prior-data conflict!

Prior-Data Conflict & Conjugate Priors

Weighted average structure is underneath all common conjugate priors for exponential family sampling distributions!

$X \overset{iid}{\sim}$ linear, canonical exponential family, i.e.

$$p(x \mid \theta) \propto \exp\left\{\langle\psi, \tau(x)\rangle - n\mathbf{b}(\psi)\right\} \qquad \Big[\psi \text{ transformation of } \theta\Big]$$

- conjugate prior: $p(\psi) \propto \exp\left\{n^{(0)}\left[\langle\psi, y^{(0)}\rangle - \mathbf{b}(\psi)\right]\right\}$
- (conjugate) posterior: $p(\psi \mid x) \propto \exp\left\{n^{(n)}\left[\langle\psi, y^{(n)}\rangle - \mathbf{b}(\psi)\right]\right\}$,

where $\quad y^{(n)} = \dfrac{n^{(0)}}{n^{(0)} + n} \cdot y^{(0)} + \dfrac{n}{n^{(0)} + n} \cdot \dfrac{\tau(x)}{n} \quad$ and $\quad n^{(n)} = n^{(0)} + n\,.$

Why Generalize Bayesian Inference?

Assigning a certain prior distribution on p

$\Longleftrightarrow$ Defining a conglomerate of probability statements (on p).

Bayesian theory lacks the ability to specify the degree of uncertainty in these probability statements.

Why Generalize Bayesian Inference?

Assigning a certain prior distribution on p

$\Longleftrightarrow$ Defining a conglomerate of probability statements (on p).

Bayesian theory lacks the ability to specify the degree of uncertainty in these probability statements.

Variance or stretch of a distribution for describing uncertainty?

Why Generalize Bayesian Inference?

Assigning a certain prior distribution on p

⟷ Defining a conglomerate of probability statements (on p).

Bayesian theory lacks the ability to specify the degree of uncertainty in these probability statements.

Variance or stretch of a distribution for describing uncertainty?

→ Does not work in the case of prior-data conflict:
In conjugate updating, the posterior variance does not depend on the degree of prior-data conflict in most cases.

Why Generalize Bayesian Inference?

Assigning a certain prior distribution on p

$\Longleftrightarrow$ Defining a conglomerate of probability statements (on p).

Bayesian theory lacks the ability to specify the degree of uncertainty in these probability statements.

Variance or stretch of a distribution for describing uncertainty?

$\Rightarrow$ Does not work in the case of prior-data conflict:
In conjugate updating, the posterior variance does not depend on the degree of prior-data conflict in most cases.

$\Rightarrow$ How to express the precision of a probability statement?

Generalized Bayesian Inference — Basic Idea

Use **set** of priors → base inferences on **set** of posteriors obtained by element-wise updating

→ numbers become intervals:

$$\mathrm{E}[p] \quad \rightarrow \quad \left[\underline{\mathrm{E}}[p],\ \overline{\mathrm{E}}[p]\right]$$

$$P(p \in A) \quad \rightarrow \quad \left[\underline{P}(p \in A),\ \overline{P}(p \in A)\right]$$

Generalized Bayesian Inference — Basic Idea

Use **set** of priors ➡ base inferences on **set** of posteriors obtained by element-wise updating ➡ numbers become intervals:

$$\mathrm{E}[p] \quad \rightarrow \quad \big[\underline{\mathrm{E}}[p],\ \overline{\mathrm{E}}[p]\big]$$

$$P(p \in A) \quad \rightarrow \quad \big[\underline{P}(p \in A),\ \overline{P}(p \in A)\big]$$

Shorter intervals ⬌ more precise probability statements

Generalized Bayesian Inference — Basic Idea

Use **set** of priors ➡ base inferences on **set** of posteriors obtained by element-wise updating ➡ numbers become intervals:

$$\mathrm{E}[p] \quad \rightarrow \quad \big[\underline{\mathrm{E}}[p], \overline{\mathrm{E}}[p]\big]$$

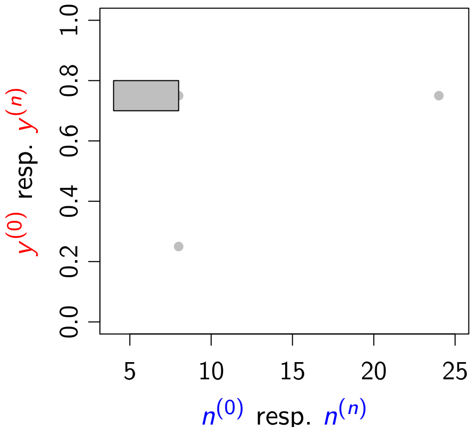
$$P(p \in A) \quad \rightarrow \quad \big[\underline{P}(p \in A), \overline{P}(p \in A)\big]$$

Shorter intervals ⬌ more precise probability statements

➡ differentiate between

- stochastic uncertainty ("risk") vs.
- non-stochastic uncertainty ("ambiguity")

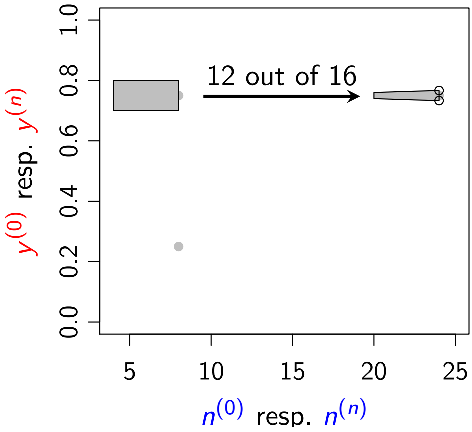
pdc-Imprecise BBM (pdc-IBBM): Walley 1991, Ch.5.4.3



no conflict:

prior $n^{(0)} \in [4, 8]$, $y^{(0)} \in [0.7, 0.8]$
data $s/n = 12/16 = 0.75$

pdc-Imprecise BBM (pdc-IBBM): Walley 1991, Ch.5.4.3

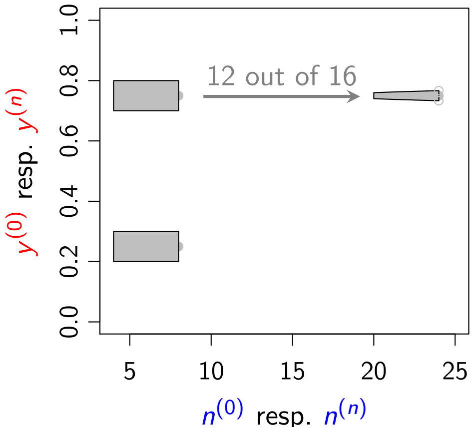


no conflict:

prior $n^{(0)} \in [4, 8]$, $y^{(0)} \in [0.7, 0.8]$
data $s/n = 12/16 = 0.75$

“spotlight” shape

pdc-Imprecise BBM (pdc-IBBM): Walley 1991, Ch.5.4.3



no conflict:

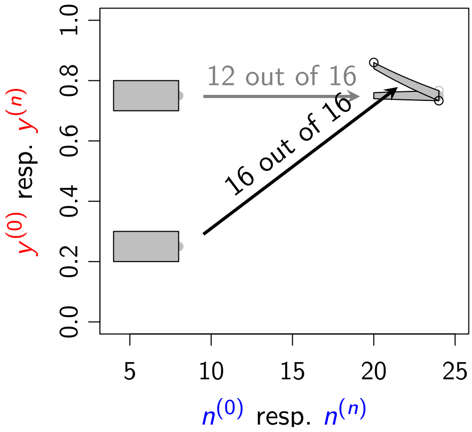
prior $n^{(0)} \in [4, 8]$, $y^{(0)} \in [0.7, 0.8]$
data $s/n = 12/16 = 0.75$

"spotlight" shape

prior-data conflict:

prior $n^{(0)} \in [4, 8]$, $y^{(0)} \in [0.2, 0.3]$
data $s/n = 16/16 = 1$

pdc-Imprecise BBM (pdc-IBBM): Walley 1991, Ch.5.4.3



no conflict:

prior $n^{(0)} \in [4, 8]$, $y^{(0)} \in [0.7, 0.8]$
data $s/n = 12/16 = 0.75$

"spotlight" shape

prior-data conflict:

prior $n^{(0)} \in [4, 8]$, $y^{(0)} \in [0.2, 0.3]$
data $s/n = 16/16 = 1$

"banana" shape

Inner Workings

- *convex* sets of distributions ("credal sets")

Inner Workings

- *convex* sets of distributions ("credal sets")
- convexity needed for consistency properties ("coherence")

Inner Workings

- *convex* sets of distributions (“credal sets”)
- convexity needed for consistency properties (“coherence”)
- sets of distributions induced by sets of parameters:
  not necessarily convex

Inner Workings

- *convex* sets of distributions ("credal sets")
- convexity needed for consistency properties ("coherence")
- sets of distributions induced by sets of parameters:
  not necessarily convex
- take convex hull of these parametric distributions:
  credal set = finite convex mixtures of parametric distributions

Inner Workings

- *convex* sets of distributions ("credal sets")
- convexity needed for consistency properties ("coherence")
- sets of distributions induced by sets of parameters: not necessarily convex
- take convex hull of these parametric distributions: credal set = finite convex mixtures of parametric distributions
- prior/posterior credal set: convex hull of distributions induced by prior/posterior parameter set

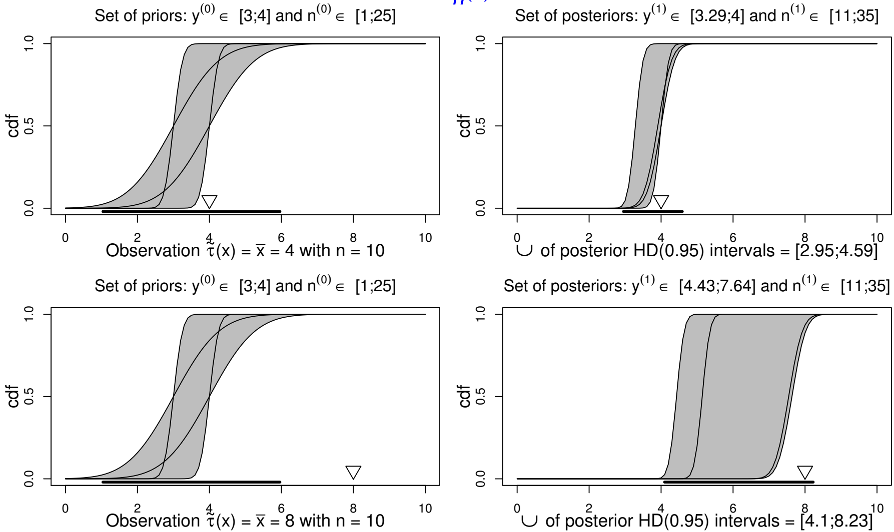
Inner Workings

- *convex* sets of distributions ("credal sets")
- convexity needed for consistency properties ("coherence")
- sets of distributions induced by sets of parameters:
  not necessarily convex
- take convex hull of these parametric distributions:
  credal set = finite convex mixtures of parametric distributions
- prior/posterior credal set: convex hull of distributions
  induced by prior/posterior parameter set
- pictures show parameter sets (that need not be convex)

Properties

Works for any canonical exponential family sampling distribution!
→ *generalized iLUCK models*, Walter & Augustin (2009)

$$X \sim \mathrm{N}(\mu, 1) \rightarrow \mu \sim \mathrm{N}\left(y^{(0)}, \frac{1}{n^{(0)}}\right)$$



Properties

Works for any canonical exponential family sampling distribution!
→ *generalized iLUCK-models*, Walter & Augustin (2009)

- $n^{(0)}$ governs precision of posterior:
  $n^{(0)} \uparrow \Longleftrightarrow$ precision$\downarrow$

Properties

Works for any canonical exponential family sampling distribution!
➡ *generalized iLUCK-models*, Walter & Augustin (2009)

- $n^{(0)}$ governs precision of posterior:
  $n^{(0)} \uparrow \Longleftrightarrow$ precision$\downarrow$
- $n \to \infty$: consistency ($y^{(n)}$ set reduces to a point at $\tau(x)/n$)

Properties

Works for any canonical exponential family sampling distribution!
➡ *generalized iLUCK-models*, Walter & Augustin (2009)

- $n^{(0)}$ governs precision of posterior:
  $n^{(0)} \uparrow \Longleftrightarrow$ precision$\downarrow$
- $n \to \infty$: consistency ($y^{(n)}$ set reduces to a point at $\tau(x)/n$)
- $y^{(0)}$ stretch $\uparrow \Longleftrightarrow y^{(n)}$ stretch $\uparrow$

Properties

Works for any canonical exponential family sampling distribution!
➡ *generalized iLUCK-models*, Walter & Augustin (2009)

- $n^{(0)}$ governs precision of posterior:
  $n^{(0)} \uparrow \Longleftrightarrow$ precision$\downarrow$
- $n \to \infty$: consistency ($y^{(n)}$ set reduces to a point at $\tau(x)/n$)
- $y^{(0)}$ stretch $\uparrow \Longleftrightarrow y^{(n)}$ stretch $\uparrow$
- inferences should be linear in posterior distributions:
  then min/max are attained at the parametric distributions
  (these are the extreme points of the credal set);
  E, Var are linear in the parametric distributions.

Properties

Works for any canonical exponential family sampling distribution!
➡ *generalized iLUCK-models*, Walter & Augustin (2009)

- $n^{(0)}$ governs precision of posterior:
  $n^{(0)} \uparrow \Longleftrightarrow$ precision$\downarrow$
- $n \to \infty$: consistency ($y^{(n)}$ set reduces to a point at $\tau(x)/n$)
- $y^{(0)}$ stretch $\uparrow \Longleftrightarrow y^{(n)}$ stretch $\uparrow$
- inferences should be linear in posterior distributions:
  then min/max are attained at the parametric distributions
  (these are the extreme points of the credal set);
  E, Var are linear in the parametric distributions.
- reaction to prior-data conflict due to different 'updating speeds' depending on $n^{(0)}$: $y^{(n)}$ moves "faster" for low $n^{(0)}$

Open Ends/Challenges

- rectangular prior set (two-dimensional interval) seems natural, but generally any shape possible

Open Ends/Challenges

- rectangular prior set (two-dimensional interval) seems natural, but generally any shape possible
- posterior parameter sets are not rectangular anyway

Open Ends/Challenges

- rectangular prior set (two-dimensional interval) seems natural, but generally any shape possible
- posterior parameter sets are not rectangular anyway
- prior shape influences the posterior inferences

Open Ends/Challenges

- rectangular prior set (two-dimensional interval) seems natural, but generally any shape possible
- posterior parameter sets are not rectangular anyway
- prior shape influences the posterior inferences
- shape can be taylored to enable desired inference properties (e.g. bonus precision if prior and data agree especially well)

Open Ends/Challenges

- rectangular prior set (two-dimensional interval) seems natural, but generally any shape possible
- posterior parameter sets are not rectangular anyway
- prior shape influences the posterior inferences
- shape can be taylored to enable desired inference properties (e.g. bonus precision if prior and data agree especially well)
- for more complex shapes, elicitation becomes more difficult

Open Ends/Challenges

- rectangular prior set (two-dimensional interval) seems natural, but generally any shape possible
- posterior parameter sets are not rectangular anyway
- prior shape influences the posterior inferences
- shape can be taylored to enable desired inference properties (e.g. bonus precision if prior and data agree especially well)
- for more complex shapes, elicitation becomes more difficult
- take two $y^{(0)}$ intervals at two different $n^{(0)}$ values?