Bayesian Approaches to Statistics and Modeling Case Study • Part I

Mark Kurzeja

Case Layout

- We are going to walk through a multi-level regression problem using a Bayesian framework
- Bayesian frameworks are be flexible enough to do variable selection, regularization, modeling of dependence, fit models where the number of parameters exceed the number of observations, and model dependence *all within the same model*
- We will barely scratch the surface of these methods in this case, but we will explore the Bayesian workflow for modeling

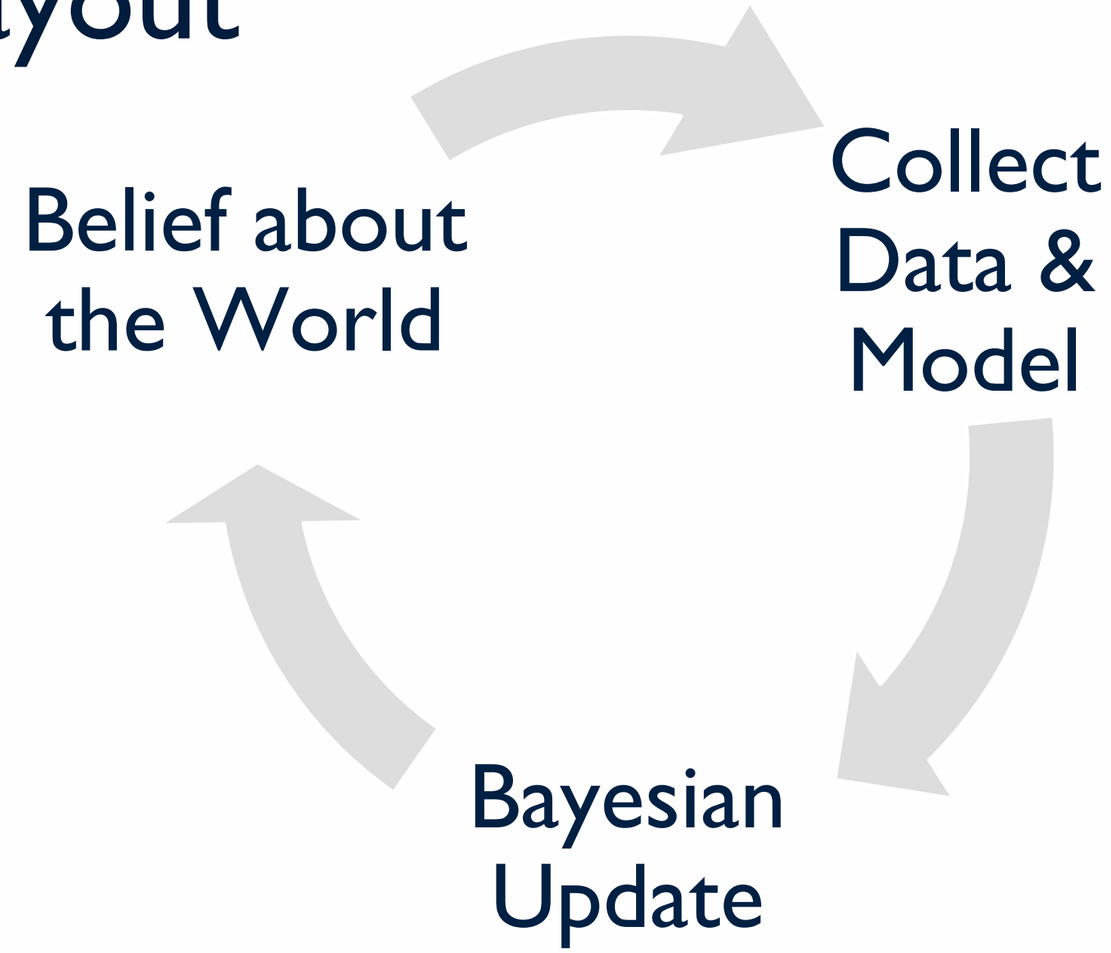
The Model

- If you're interested in further exploration of Bayesian modeling **Doing Bayesian Data Analysis** by John Kruschke or **Bayesian Data Analysis** by Andrew Gelman
- Doing Bayesian Data Analysis is a great introduction to Bayesian analysis and is one of the most approachable books on the subject
- Bayesian Data Analysis is a great graduate-level textbook on the subject and written by some of the STAN team

Case Layout

In order to proceed, we need to remember the steps to a Bayesian approach

- **Belief:** Establish a belief about the world
  - Includes Prior and Likelihood functions
- **Model:** Use data and probability, in accordance with your belief, to update your model
  - Check that your model agrees with your original data
- **Update:** Update your view of the world based on your model



Case Layout

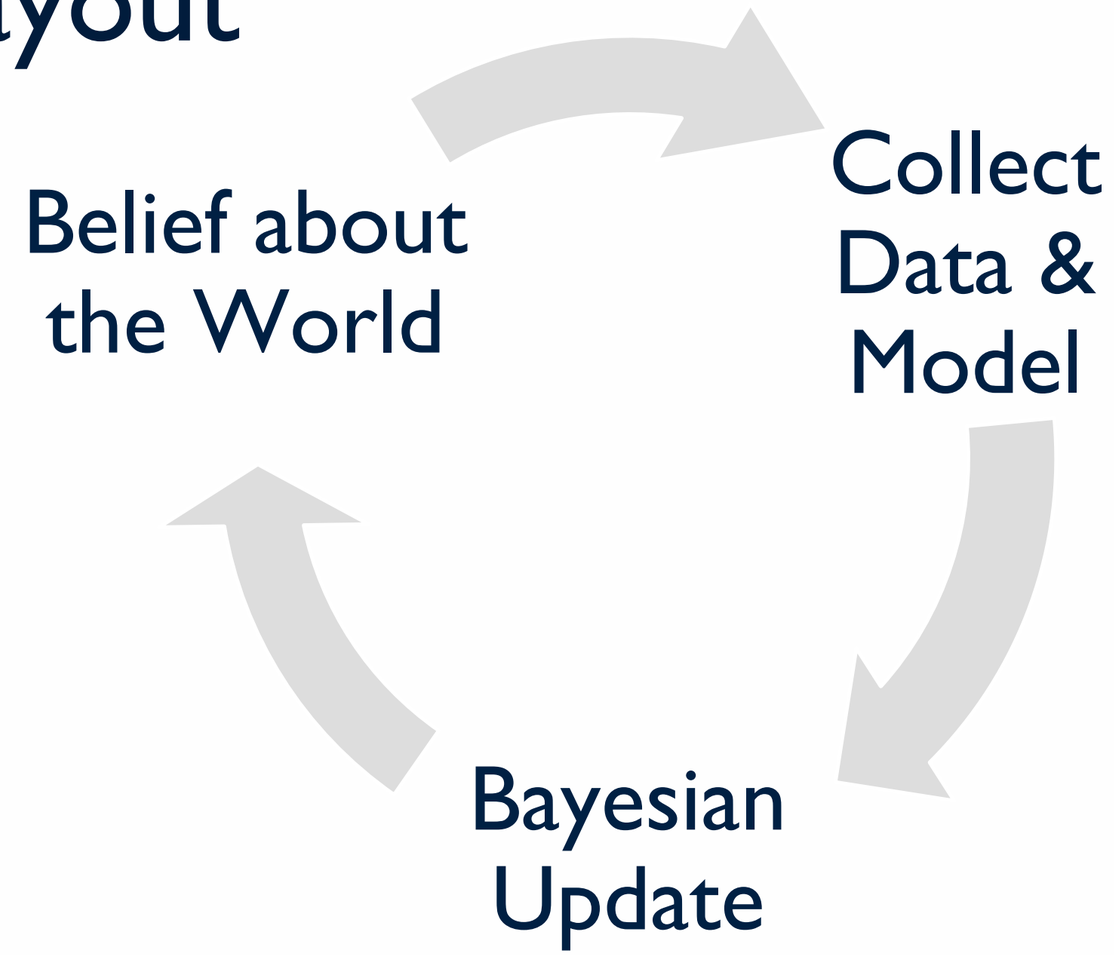
In order to proceed, we need to remember the steps to a Bayesian approach

- **Belief:** Establish a belief about the world
  - Includes Prior and Likelihood functions
- **Model:** Use data and probability, in accordance with your belief, to update your model
  - Check that your model agrees with your original data
- **Update:** Update your view of the world based on your model

Case Layout

In order to proceed, we need to remember the steps to a Bayesian approach

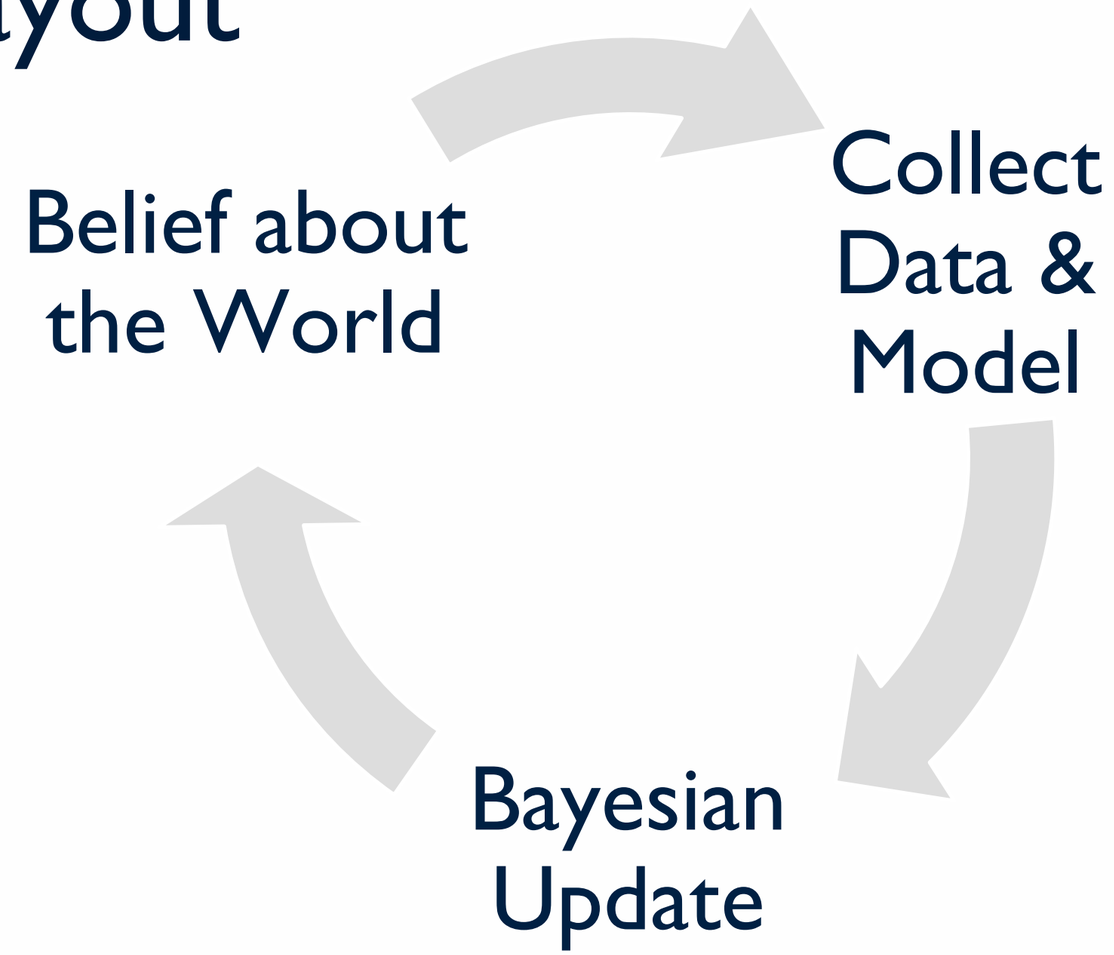
- **Belief:** Establish a belief about the world
  - Includes Prior and Likelihood functions
- **Model:** Use data and probability, in accordance with your belief, to update your model
  - Check that your model agrees with your original data
- **Update:** Update your view of the world based on your model



Case Layout

In order to proceed, we need to remember the steps to a Bayesian approach

- **Belief:** Establish a belief about the world
  - Includes Prior and Likelihood functions
- **Model:** Use data and probability, in accordance with your belief, to update your model
  - Check that your model agrees with your original data
- **Update:** Update your view of the world based on your model



The Data

- **Question:** Does a child's IQ have a relationship with the IQ of their mother?
- **Data:** National Longitudinal Survey of Youth. Source: Gelman and Hill (2007)
- 434 observations

	kid_score	mom_hs	mom_iq	mom_age
1	65	1	121	27
2	98	1	89	25
3	85	1	115	27
4	83	1	99	25
5	115	1	93	27
6	98	0	108	18

The Model

- We will use a linear model
- We will start out with the most basic of linear regression models

Regression Form

$$ChildIQ = \beta_0 + \beta_1(momIQ) + \beta_2(momAge)$$

Belief about the World → Collect Data → Bayesian Update

The Data

- Up until now, we haven't done anything different than what we have done in the past
- The Bayesian framework, however, says that we have to specify prior distributions for our beliefs as well as likelihoods

Key Point: Every parameter must begin with a distribution that captures our beliefs. We call these **priors.**

The Model

- In a Bayesian setting, we need priors on each of our parameters
- $\beta_0 \sim N(0,20)$ I put a relatively weak prior on the intercept term to allow it to vary a lot if it needs to. The data will be centered and so this isn't a large concern
- $\beta_1 \sim N(1,5)$ I put a prior centered at one with a wide variance to account for the fact that I expect that a child's IQ is able to be predicted very well by a mother's IQ but I'm not certain
- $\beta_2 \sim N(0,5)$ I put a prior centered at zero because I have no idea how age will affect IQ. The large variance accounts for my ignorance in what this value may be

$$ChildIQ_i \sim N(\beta_0 + \beta_1(momIQ_i) + \beta_2(momAge_i), \sigma_{error})$$

Belief about the World → Collect Data → Bayesian Update

The Model

- Two different people may use two different priors and conclude two different things (!)
- If this model seems subjective, it is!! That is the point! My assumptions are mine and are brought to the forefront of the analysis. I am not an expert in this area and so I express that in the uncertainty that I put on my estimates!
- I could ask an expert for their opinion to get better priors or use previous studies to inform my priors.

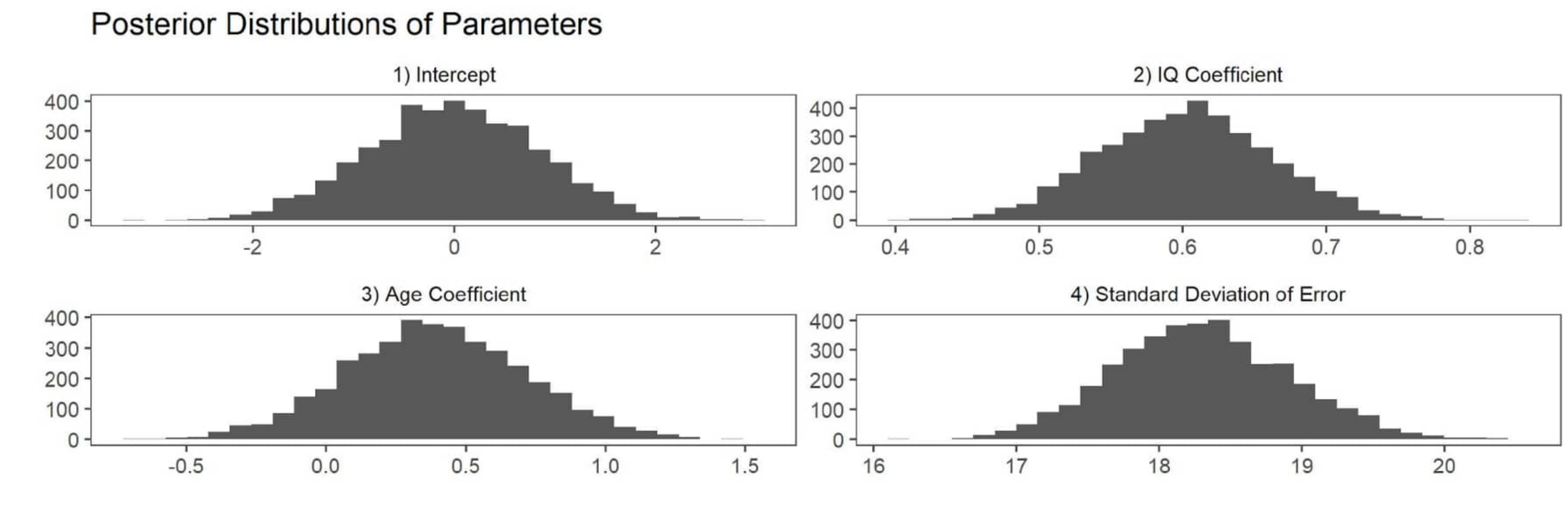
$$ChildIQ_i \sim N(\beta_0 + \beta_1(momIQ_i) + \beta_2(momAge_i), \sigma_{error})$$

Belief about the World → Collect Data → Bayesian Update

The Model

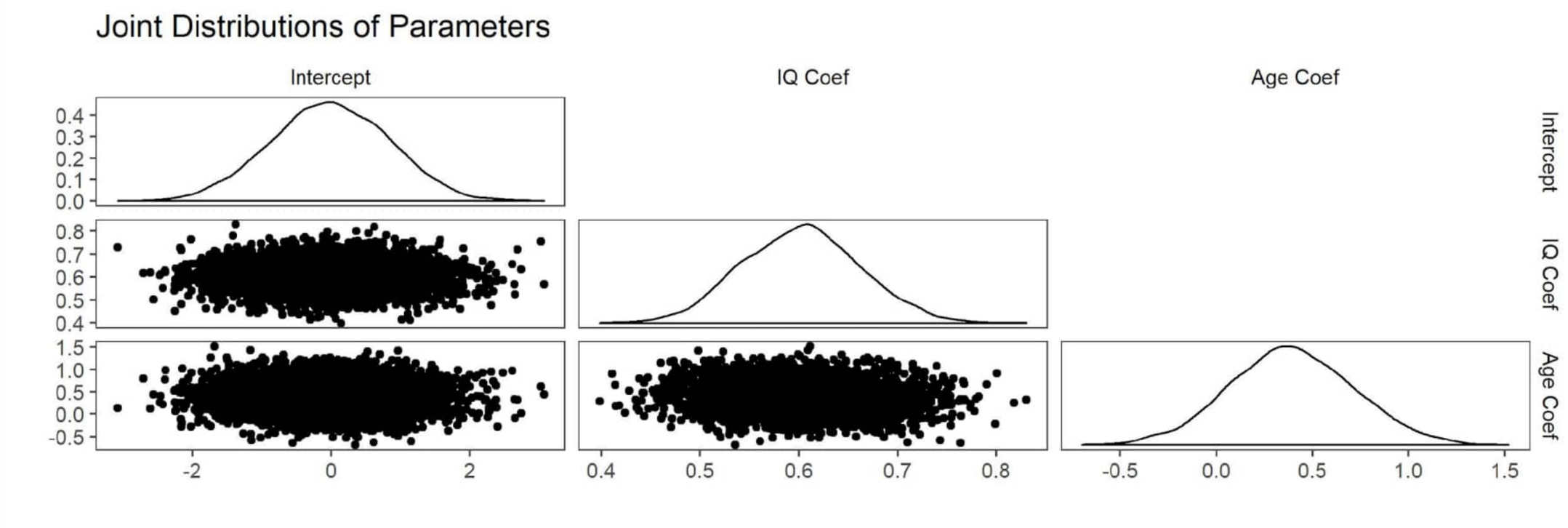
- I fit these models using a program called STAN, a domain specific language for Bayesian modeling
- For some problems, like the simple regression problem, the Bayesian models are tractable. However, they quickly become intractable as they are expanded and need to use computationally intensive sampling methods to fit them
- The data was centered before fitting

The Model – The Posteriors



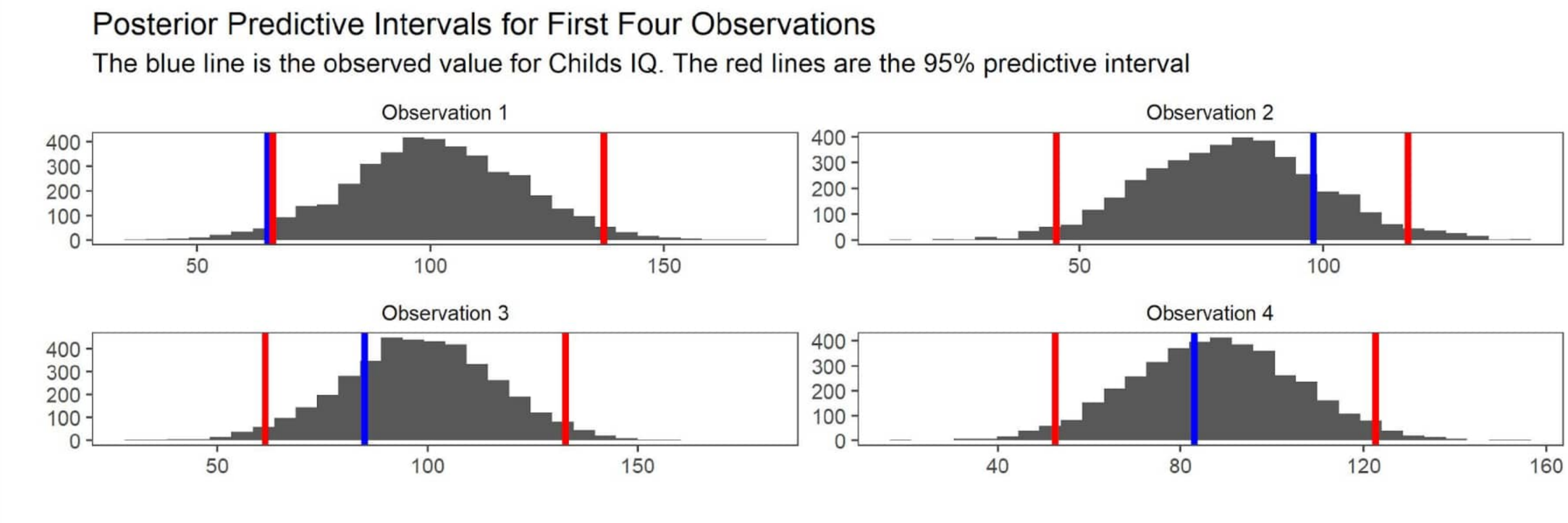
Belief about the World → Collect Data → Bayesian Update

The Model – The Posteriors



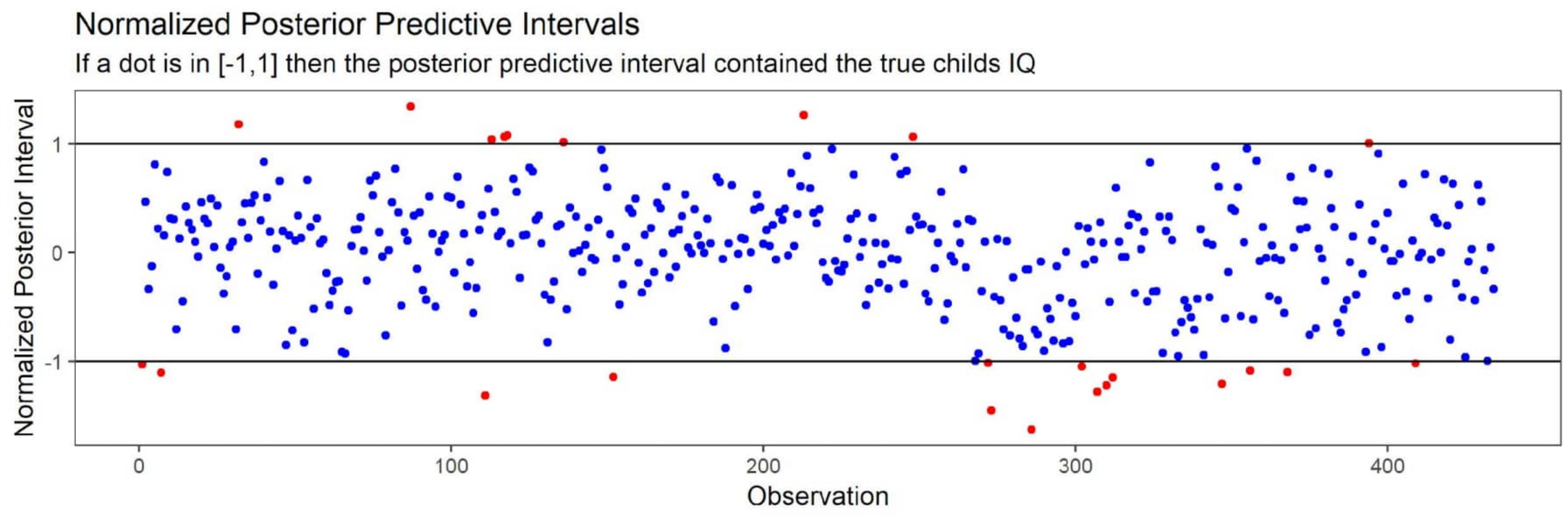
Belief about the World → Collect Data → Bayesian Update

The Model – The Posteriors



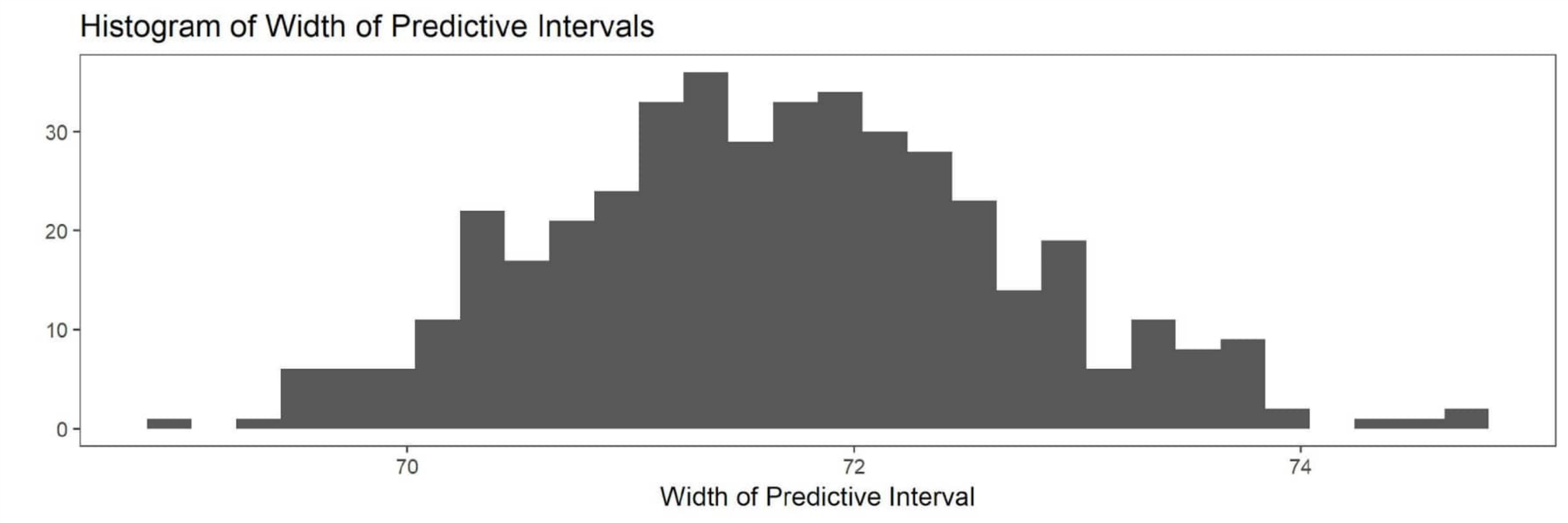
Belief about the World → Collect Data → Bayesian Update

The Model – The Posteriors



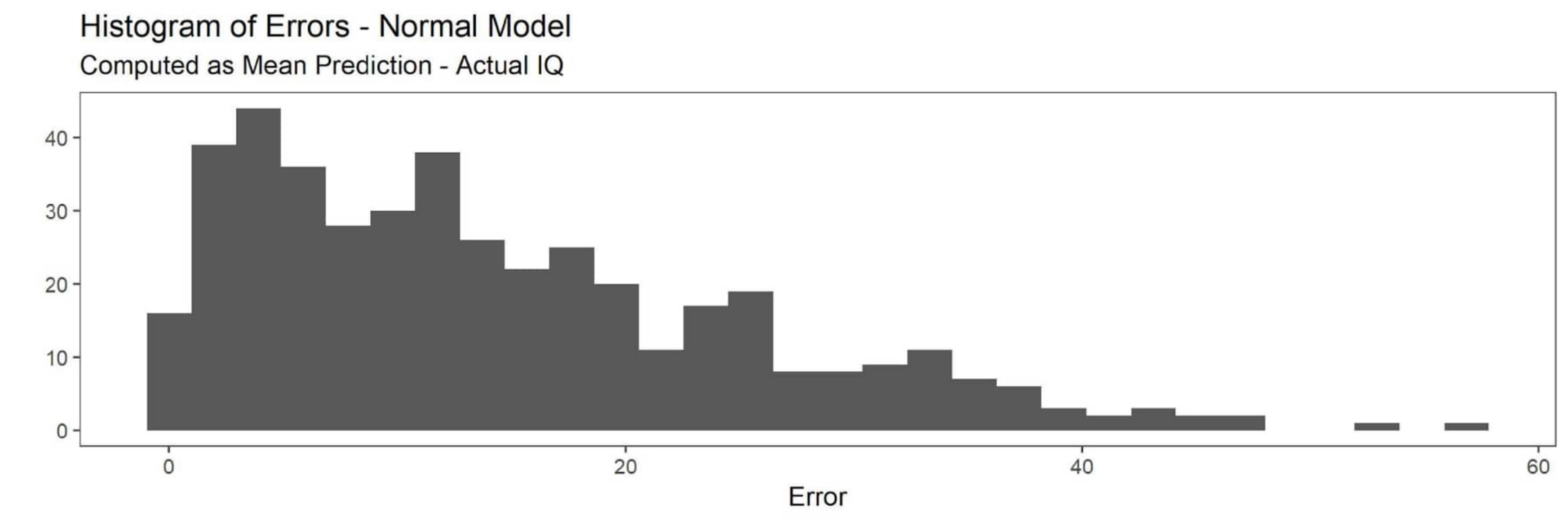
Belief about the World → Collect Data → Bayesian Update

The Model – The Posteriors



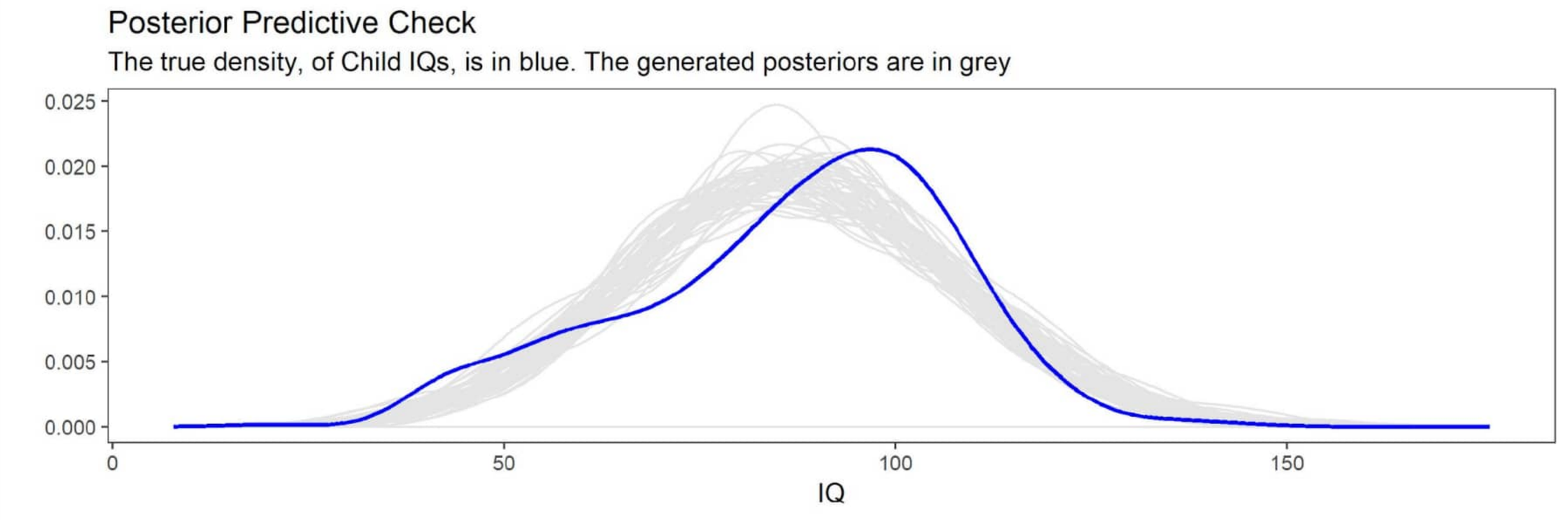
Belief about the World → Collect Data → Bayesian Update

The Model – The Posteriors



Belief about the World → Collect Data → Bayesian Update

The Model – The Posteriors



Belief about the World → Collect Data → Bayesian Update

The Model - Observations

- We have not accounted for the high school education variable
- We have a reasonable number of observations that fall outside of their 95% predictive intervals ($\approx 5\%$)
- The width of the predictive intervals is pretty large. We should see if we can do better
- The estimates systematically overestimate IQ's of around 65-80 and underestimate those around 90-115

Belief about the World → Collect Data → Bayesian Update

The Model - Observations

- Next time, we are going to explore updates that we can make to this model
- We will look to incorporating the high school status of the mother and working with a hierarchal structure will improve our model
- We will need to come up with some way of addressing the skew of the data