Big ideas in data science

Laplace's big idea (1810). Central Limit Theorem. The sum of many individual stochastic processes, each of which could come from any of a variety of distributions, tends to a normal distribution. (Extending De Moivre who showed the same for binomial processes in 1733). McElreath Ch 4.

Main messages from McElreath Ch 4

- Language for describing models

e.g. a linear model

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 x_i \\
\beta_0 &\sim \text{Normal}(178, 100) \\
\beta_1 &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Uniform}(0, 50)
\end{aligned}
$$

Mean model

In the book:

$$y_i \sim \text{Normal}(\mu, \sigma)$$
$$\mu \sim \text{Normal}(178{,}100)$$
$$\sigma \sim \text{Uniform}(0{,}50)$$

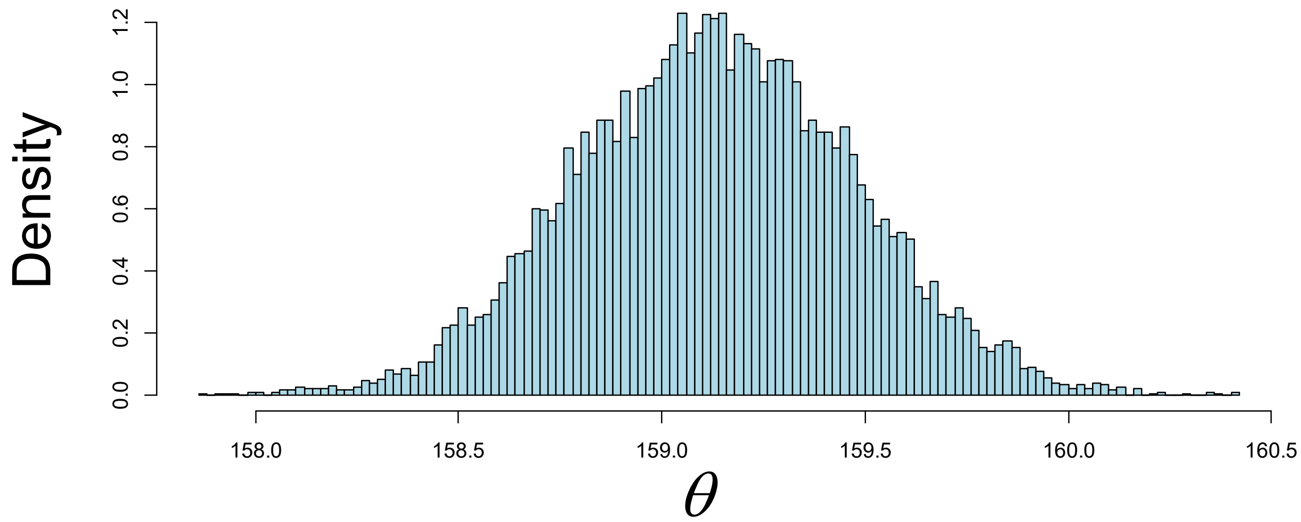
Alternative:

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \beta_0$$
$$\beta_0 \sim \text{Normal}(178{,}100)$$
$$\sigma \sim \text{Uniform}(0{,}50)$$

Thus, the mean is a special case of the linear model

Main messages from McElreath Ch 4

- Histogram is the posterior distribution



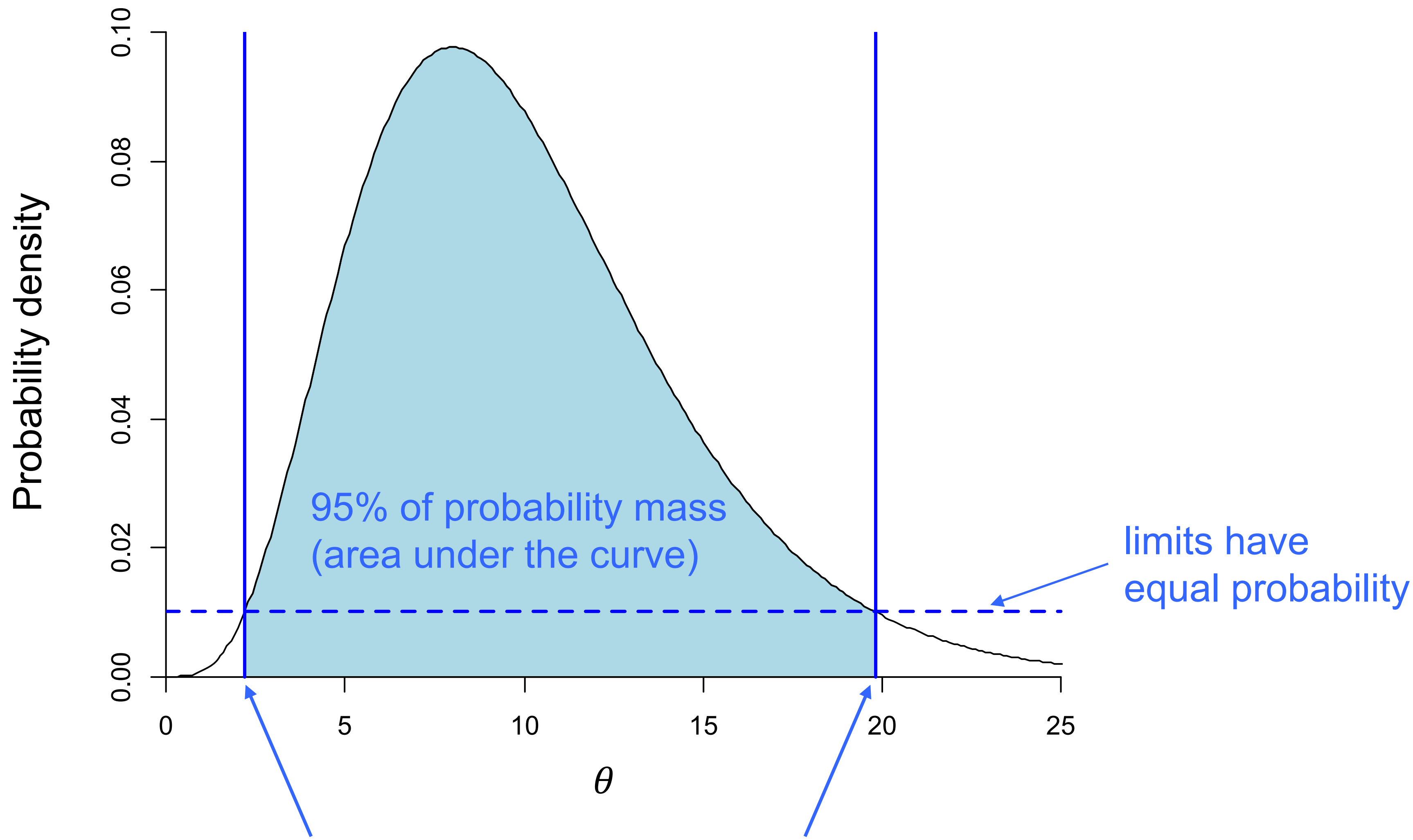
- Obtain all our inferences (means, credible intervals, prediction intervals) from the posterior samples of parameters

Main messages from McElreath Ch 4

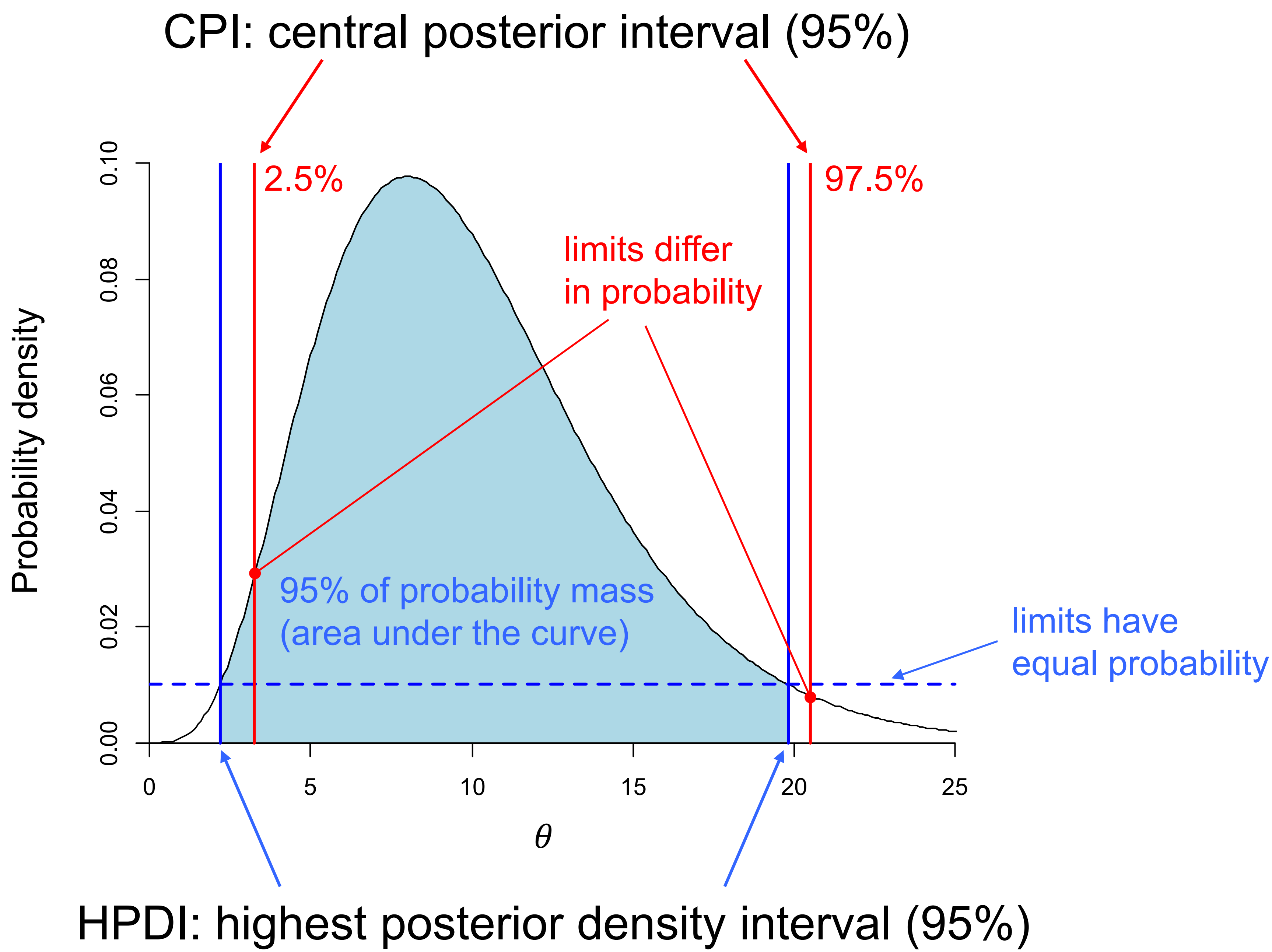
- Credible intervals are probabilities for model quantities (e.g. parameters or the average relationship)

Credible intervals

- Plausibility intervals
- HPDI vs CPI



HPDI: highest posterior density interval (95%)



CPI: central posterior interval (95%)
2.5%
97.5%
limits differ
in probability
95% of probability mass
(area under the curve)
limits have
equal probability
Probability density
0.00
0.02
0.04
0.06
0.08
0.10
0
5
10
15
20
25
θ
HPDI: highest posterior density interval (95%)

Main messages from McElreath Ch 4

- Credible intervals are probabilities for model quantities (e.g. parameters or the average relationship)
- Prediction intervals are probabilities for a new data point

  Uncertainty in parameters + uncertainty in the data generating process

Main messages from McElreath Ch 4

- Credible intervals are probabilities for model quantities (e.g. parameters or the average relationship)
- Prediction intervals are probabilities for a new data point
  - Uncertainty in parameters + uncertainty in the data generating process
- What do you think about 89% intervals?

Main messages from McElreath Ch 4

- You can derive quantities (e.g. height at weight 50 kg) from the posterior samples
- Correlation among parameters can be obtained from the posterior samples
- Flat priors correspond to non-Bayesian approaches. Useful but there often better alternatives

Main messages from McElreath Ch 4

- Polynomials: nonlinear models that are linear in the parameters
- Centering can help the training algorithm converge
- Standardizing allows you to compare predictors on a common scale

My recommendations: points of departure

- Plot histograms instead of densities
- Credible intervals usually make the most sense (most coherent) compared to central posterior intervals
  - calculate HPDI
  - unless CPI is more stable estimate of HPDI
- Don't use the quadratic approximation (also called Laplacian approximation)
  - always use MCMC

4H1. The weights listed below were recorded in the !Kung census, but heights were not recorded for these individuals. Provide predicted heights and 89% intervals (HPDI) for each of these individuals. That is, fill in the table below, using model-based predictions.

Individual	weight	expected height	89% interval
1	46.95		
2	43.72		
3	64.78		
4	32.59		
5	54.63		