Normal model; summaries; sampling from the posterior

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

September 1, 2021

Outline

Last time:

- Exploring choice of priors
- Approximate methods for inference
- A little on summaries
- Intro to a normal-likelihood model

Today:

- Continuing the normal model
- Applications of random sampling

Note: refer to Ch. 3 of *Rethinking*

Return to the normal model

The normal model, known variance

Example we talked about last time:

- Target for inference: average combined score in NCAA tournament men's basketball games
- Source data: scores from all games 1960 – 1995
- Model:

$$y_i \sim \text{Normal}(\theta, \sigma)$$
$$\theta \sim \text{Normal}(\mu_0, \tau_0)$$

Last time we picked $\mu_0 = 200, \tau_0 = 60$

Calculating the posterior

Assume we start with one observation y. Since we are using a conjugate prior, the posterior is analytically expressible:

$$
\begin{aligned}
p(\theta|y) &\propto \exp\left(-\frac{1}{2}\frac{(y-\theta)^2}{\sigma^2}\right)\exp\left(-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\tau_0^2}\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right) \\
&= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}\right)\theta^2-\left(\frac{2y}{\sigma^2}+\frac{2\mu_0}{\tau_0^2}\right)\theta+\frac{y^2}{\sigma^2}+\frac{\mu_0^2}{\tau_0^2}\right)
\end{aligned}
$$

Then some algebra happens...

Calculating the posterior

$$\theta|y \sim \text{Normal}(\mu_1, \tau_1)$$

where

$$\mu_1 = \frac{\frac{1}{\sigma^2}y + \frac{1}{\tau_0^2}\mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}$$

$$\frac{1}{\tau_1} = \frac{1}{\sigma_2} + \frac{1}{\tau_0^2}$$

The inverse variances $1/\sigma^2, 1/\tau^2$ are called the *precisions* of these distributions

(What's the algebra? Complete the square (exercise 2.14(a)) in BDA3)

The posterior as a compromise

Three ways of writing the posterior mean of θ:

$$\mu_1 = \frac{\frac{1}{\sigma^2} y + \frac{1}{\tau_0^2} \mu_0}{\frac{1}{\sigma^2} + \frac{1}{\tau_0^2}}$$

$$\mu_1 = \mu_0 + (y - \mu_0) \frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

$$\mu_1 = y - (y - \mu_0) \frac{\sigma^2}{\sigma^2 + \tau_0^2}$$

- Weighted average of μ_0 and y
- Prior mean μ_0 adjusted toward the data
- Data "shrunk" toward the prior mean

In all: the weight of μ_0 is determined by the prior precision τ_0^{-2}

Generalizing to many observations

We don't have to iterate this process a thousand times to incorporate our thousand games (although the ability to incorporate observations one by one can be considered a feature of the Bayesian approach); the posterior depends on $y_1, y_2, \ldots$ only through the sample mean $\bar{y}$[1]

$$\theta|(y_1, y_2, \ldots y_n) \sim \text{Normal}(\mu_n, \tau_n)$$

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

$$\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

[1] $\bar{y}$ is called a *sufficient statistic* in this model

Normal model, unknown variance

Adding a parameter

The assumption of known variance doesn't make a ton of sense in practice. So let's add σ back in as a free parameter.

- We should now think of the model as specifying a joint probability distribution on (μ, σ)
- The likelihood depends on each parameter:

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

- Inference target is the joint posterior

$$p(\mu,\sigma|y) \propto p(y|\mu,\sigma)p(\mu,\sigma)$$

Choosing priors

The prior now is really a joint distribution on μ, σ

- Simpler to specify independent priors for each
- Non-independence can still come through in the posterior

BDA3 makes an argument for a prior on σ which is proportional to $(\sigma^2)^{-1}$, especially in tandem with a flat prior on μ

In the following, I'll use:

$$\sigma \sim \text{Exponential}(1/10)$$

Posterior distribution

With this combination of priors, we can't easily write down the posterior distribution. (There are choices of prior, conjugate and non-conjugate, where you can.)

So: fall back to quadratic approximation.

Quadratic approximation results:

	mean	sd	hdi_3%	hdi_97%
mu	145.074	2.297	140.753	149.395
sigma	23.036	1.602	20.024	26.049

Point estimate; sd of the posterior for that variable; posterior interval

Summarizing and displaying inferences

Intervals

- Fixed-value intervals
  - estimate probability that a parameter value lies in a certain range
- Fixed-probability intervals
  - central interval
  - highest-density interval

Central intervals, by any other name

Common terminology:

- Confidence interval (common in frequentist statistics)
- Credible interval (common in Bayesian statistics)

McElreath's book: *compatibility interval* – values *compatible* with the model, conditional on the data

Point estimates

Three common point estimates:

- posterior mode aka MAP (maximum a posteriori)
- posterior mean/expectation
- posterior median

Choosing:

- MAP estimate often the least reliable
- minimizing a loss function

Making inferences from a sample

Sampling from the posterior is a useful way to produce at least some intervals:

- Generate random samples from the posterior distribution
- For a central interval: clip off the top/bottom

and point estimates:

- Posterior mean/median: just take a sample mean/median
- Posterior mode: estimate a density from the sample

Sampling methods

Drawing random samples:

- Grid approximation:
  - Normalize the posterior and draw from the grid
  - Python: `np.random.choice`; R: `sample`
- Quadratic approximation:
  - Assumed posterior multivariate normal; draw from there
  - `sp.stats.multivariate_normal`
- Markov chain Monte Carlo
  - Works with arbitrary posterior, no normalization required

Using posterior samples for visualization

Posterior samples can also be useful for visualizing the results of a model, in a way that a summary table doesn't quite capture

What does a single sample from our posterior distribution represent?

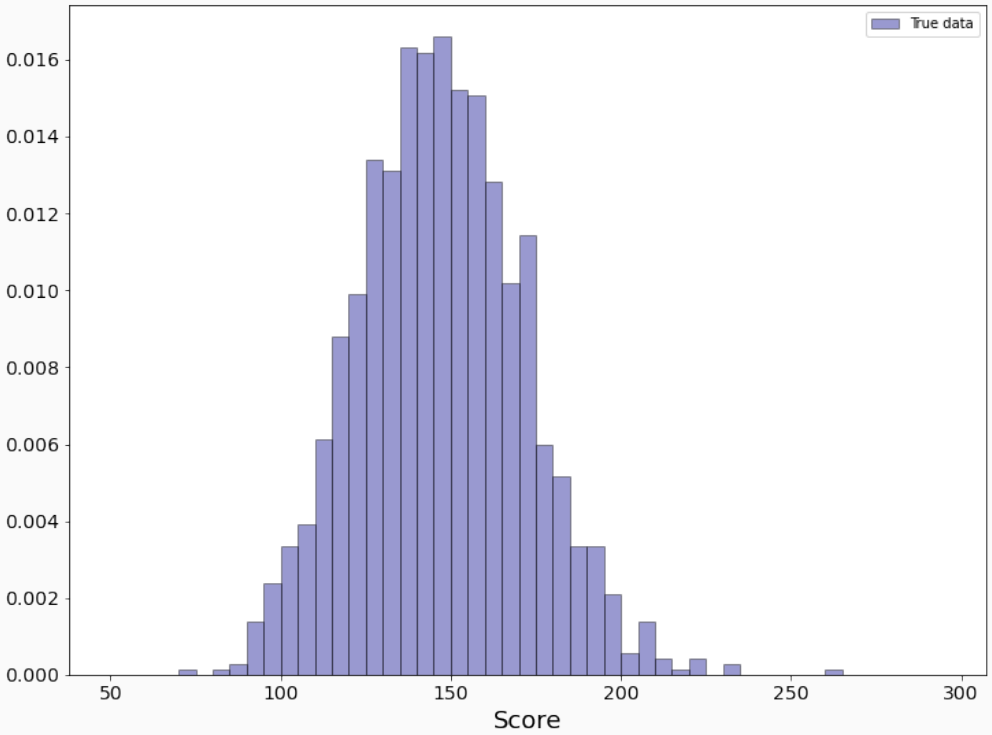
- (μ, σ) – a single pair of mean & SD for the observation model

$$y \sim \text{Normal}(\mu, \sigma)$$

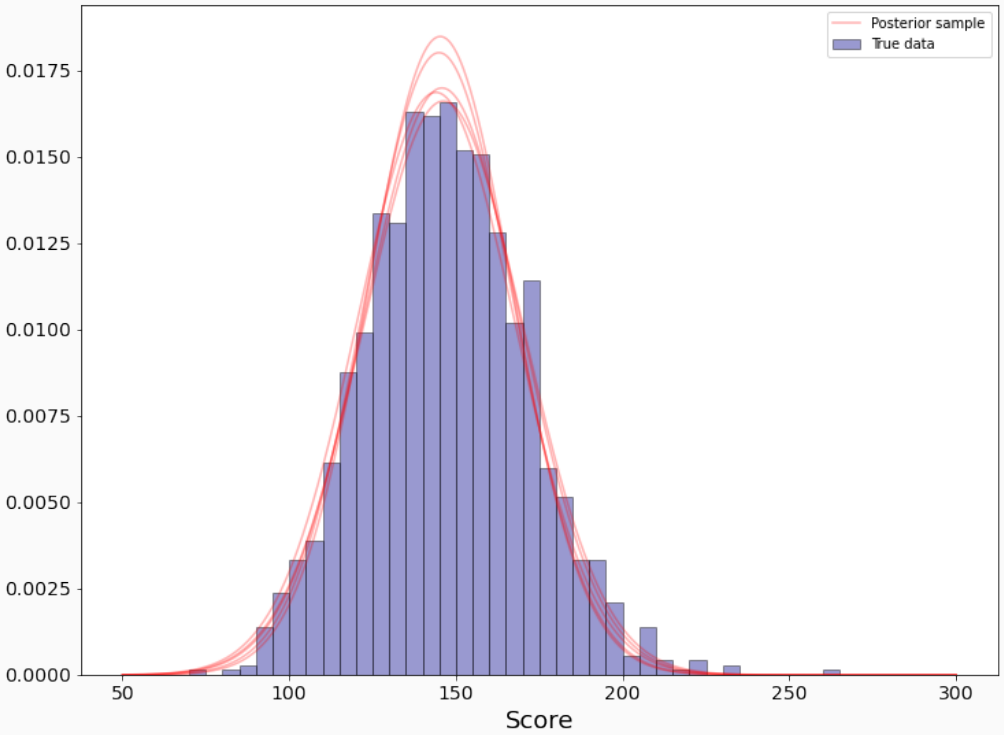
- The posterior is full of normal curves – it's a distribution over the space of possible normal distributions

To see this, draw a few curves from the hat.

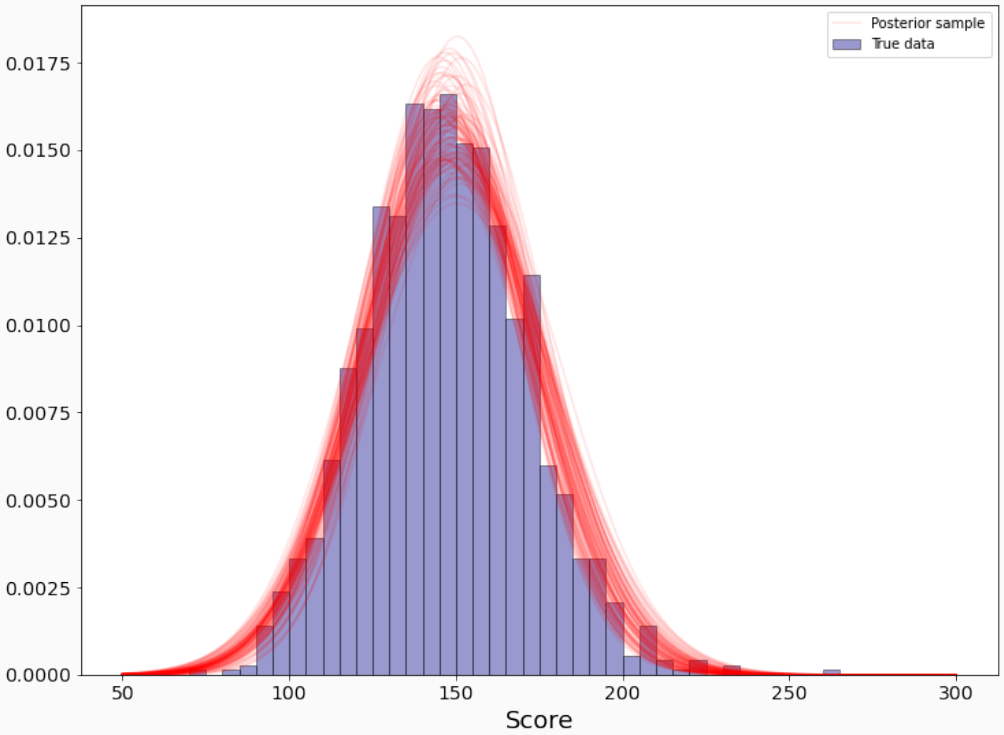
A histogram of the real data



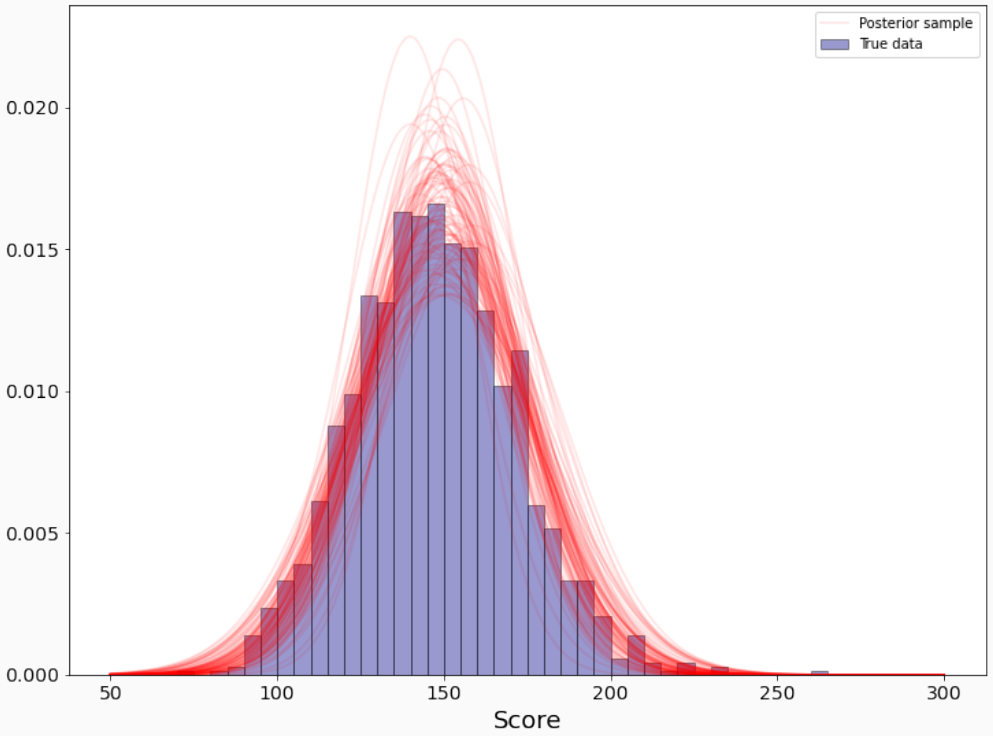
Put a curve on it



Put a curve on it

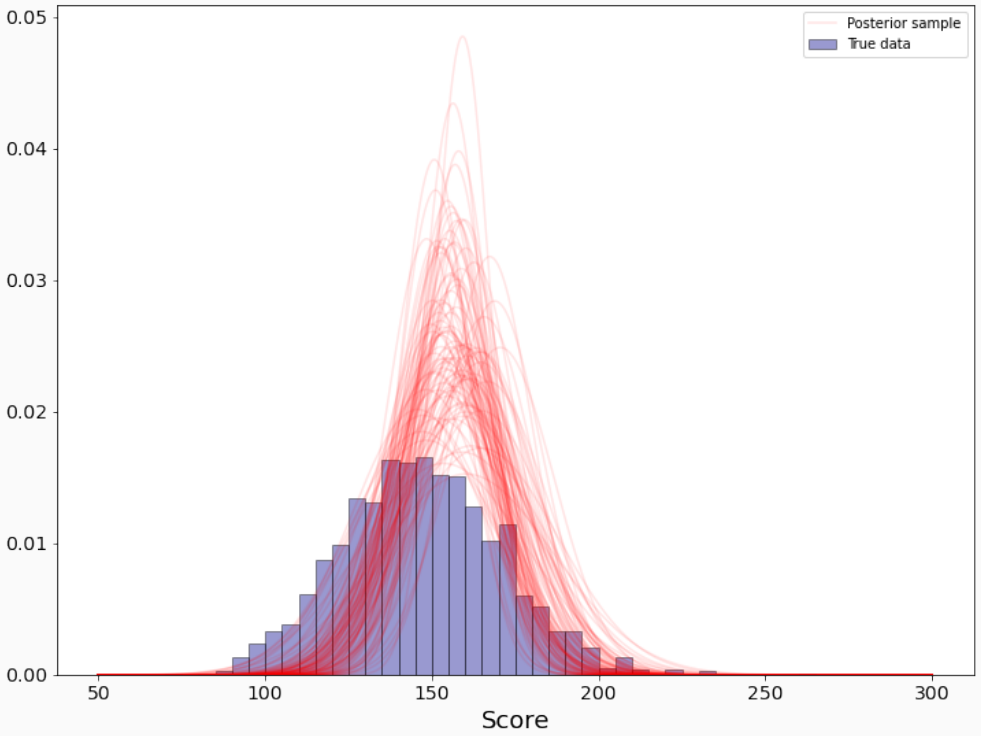


Put a curve on it



Inference from only 40 games

Put a curve on it



Inference from only 10 games

Comparing inferences from two methods

Posterior inference for the basketball model: quadratic approximation vs. MCMC

Quadratic:

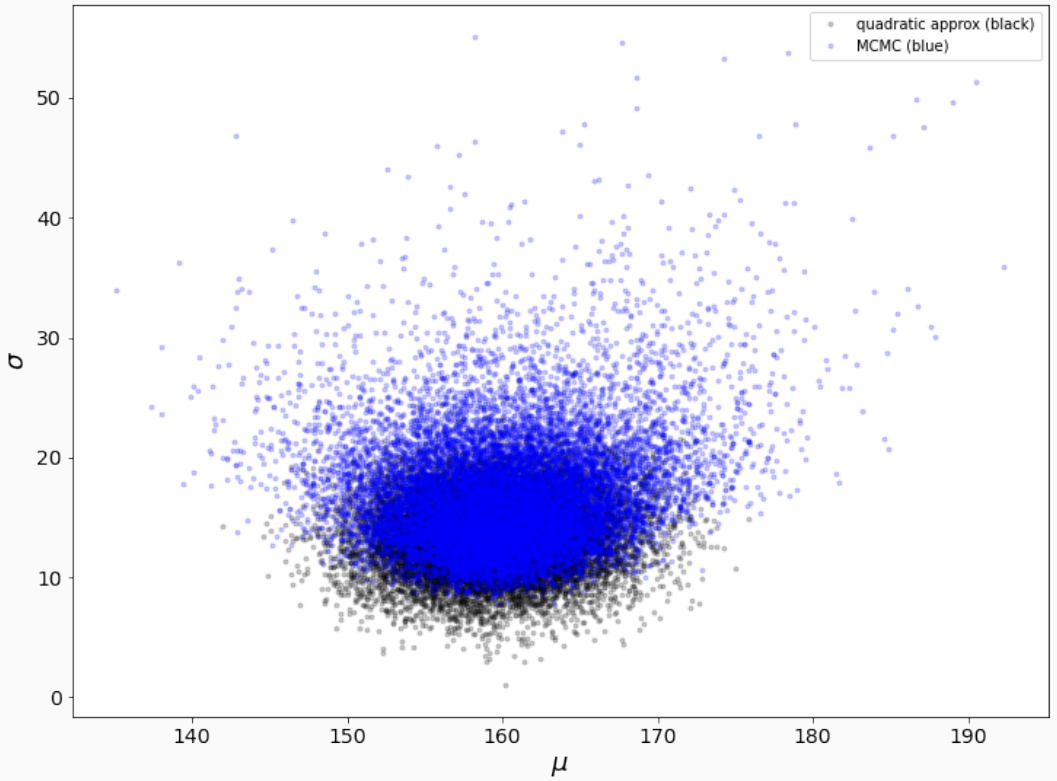
	mean	sd	hdi_3%	hdi_97%
mu	145.074	2.297	140.753	149.395
sigma	23.036	1.602	20.024	26.049

MCMC:

	mean	sd	hdi_3%	hdi_97%
mu	145.099	2.408	140.524	149.560
sigma	23.444	1.676	20.388	26.641

Comparing inferences from two methods

Posterior inference for the basketball model: quadratic approximation vs. MCMC



Comparing inferences from two methods

What's different between the two methods?

- σ has a longer tail in the MCMC sample
- μ more spread out when σ is larger
- μ, σ not really independent in the posterior

Predictive sampling

Simulating future observations

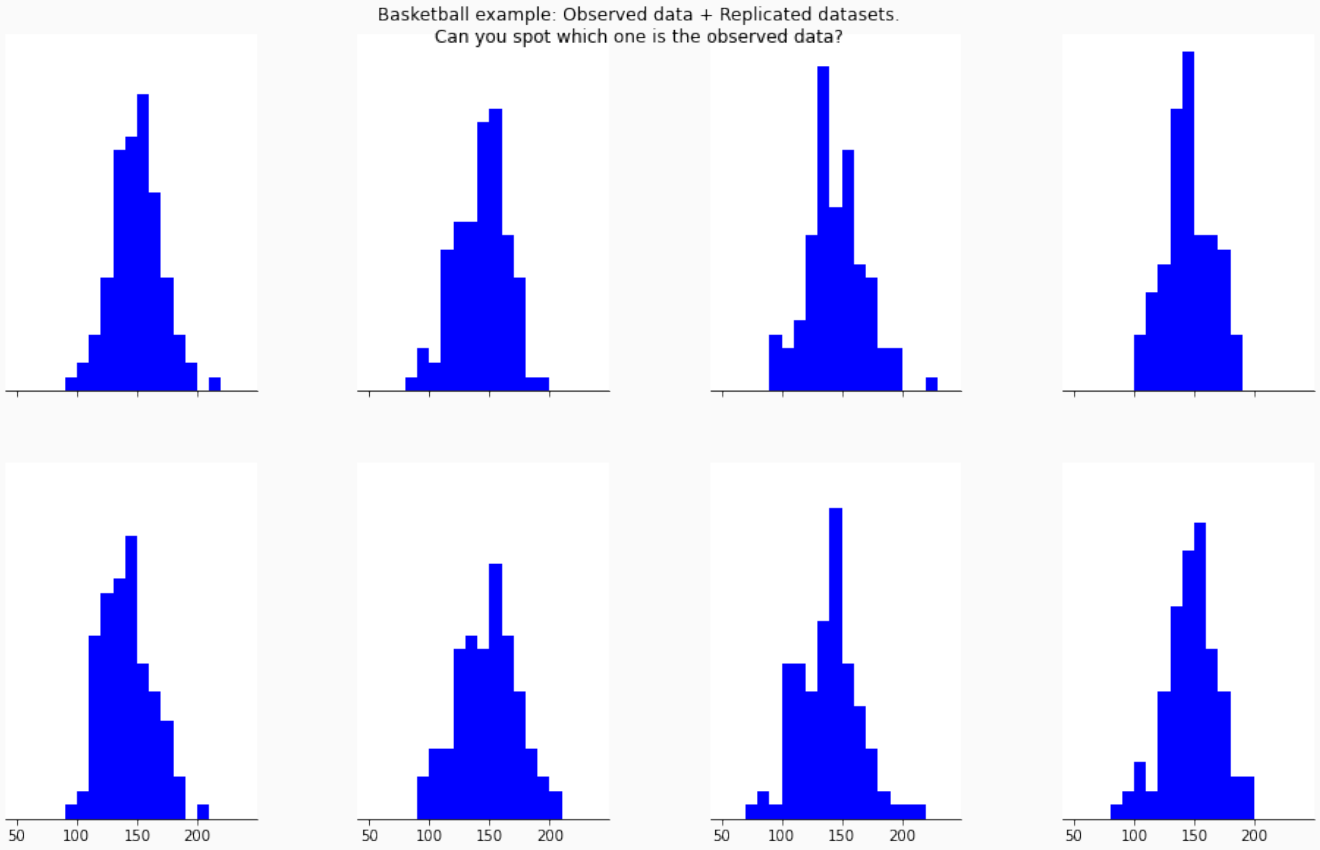
- Sampling from the posterior – produce plausible values of (μ, σ)
- Since the models are generative, we can also produce predictions of y
- Process:
  - Draw a pair $(\hat{\mu}, \hat{\sigma})$ from the posterior
  - Draw a value $y \sim N(\mu, \sigma)$

These samples come from the *posterior predictive distribution*

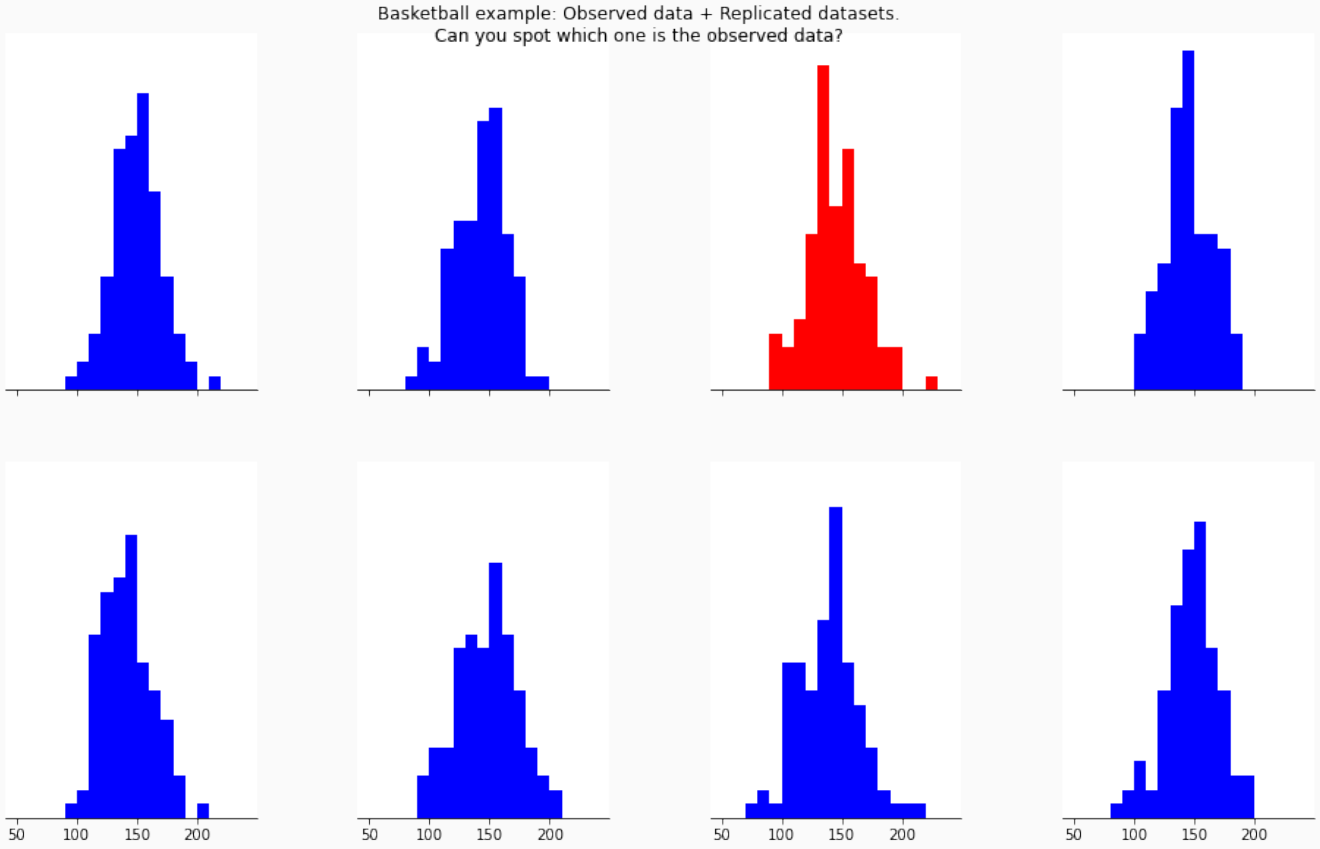
Why sample from the predictive distribution?

- Forecasting: application of models, especially in ML contexts, often involves
- Model evaluation: a good model should be able to produce simulated data that "resembles" real data (recall what we did with the kidney cancer example)
- Software testing: use predictive sampling with known models / fake data to ensure consistency
- Model design and prior evaluation: help understand structure of model and implications of the prior; power analysis

Checking the model by posterior predictive sampling



Checking the model by posterior predictive sampling



Checking our prior: prior predictive simulations

Prior predictive simulations: draw observations (i.e. values of y_i) using the prior distribution

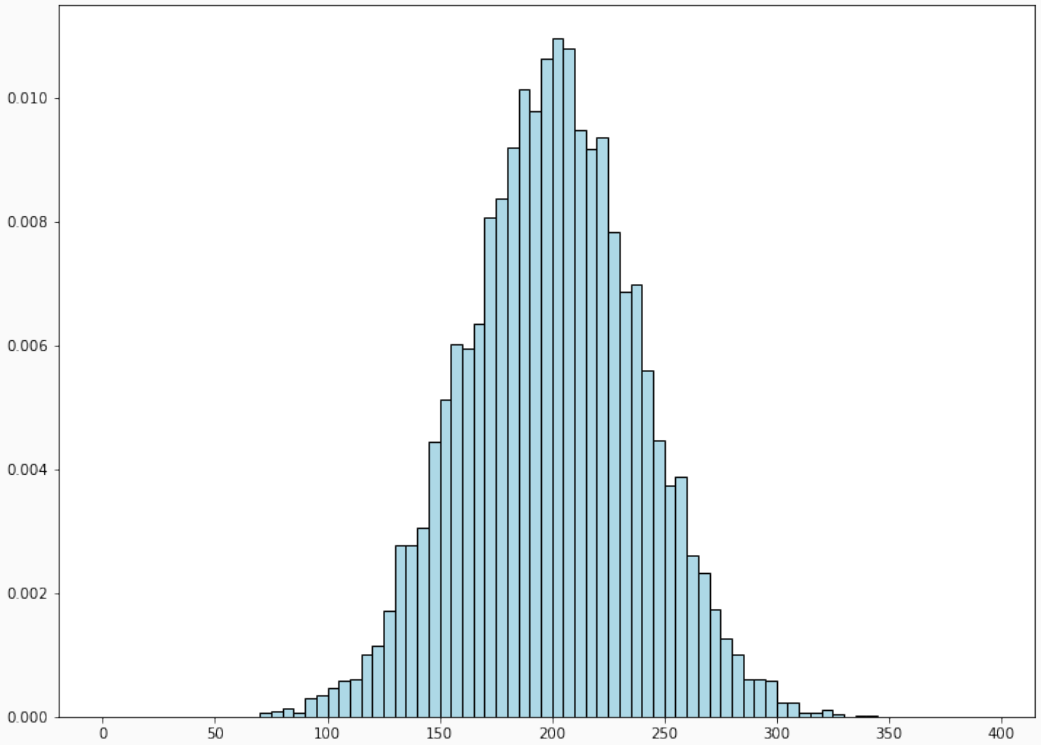
This can be used to check the reasonableness of a prior, by making sure it doesn't produce impossible results.

- We're not looking for the prior predictions to be a perfect model for the data
- But, if our predictive draws have games with negative score, or teams scoring 500 points, maybe something is off

Procedure:

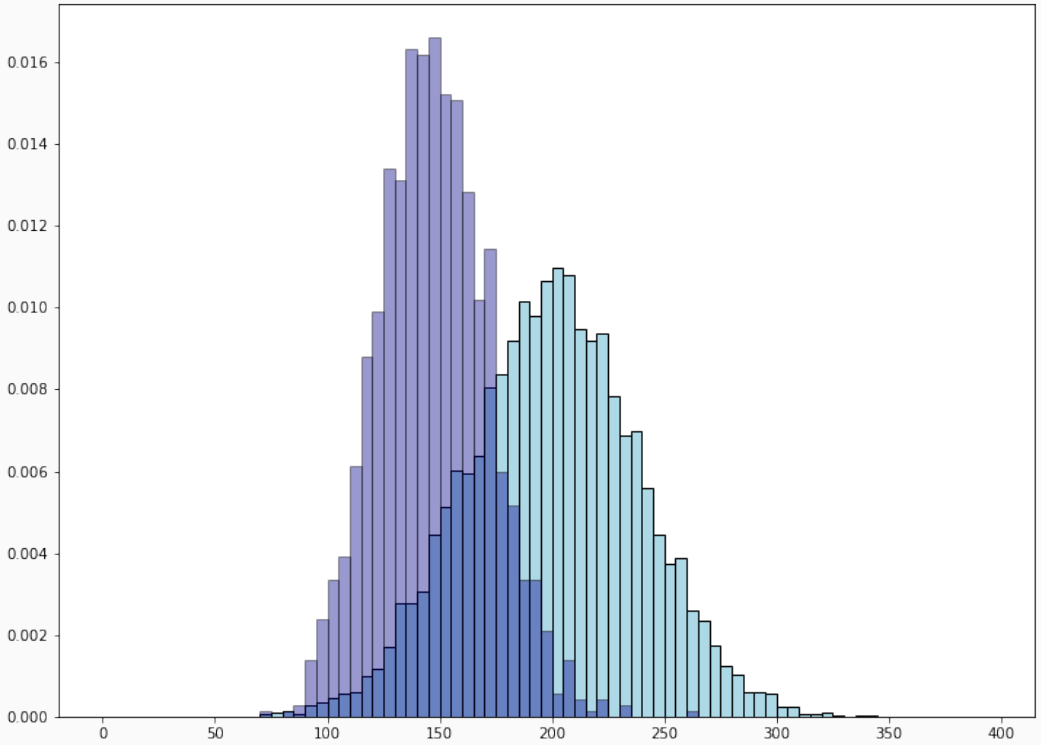
- draw n samples of θ from the prior $\mathrm{Normal}(\mu_0, \tau_0)$
- draw one sample of y from the likelihood $\mathrm{Normal}(\theta, \sigma)$

Checking our prior: prior predictive simulations



Do these predictions resemble the real data?

Checking our prior: prior predictive simulations



Do these predictions resemble the real data? No (and they shouldn't)

Informative vs. uninformative priors

Most often, priors are categorized as *informative* or *uninformative priors* depending on whether they incorporate outside scientific information

- informative priors: bring in knowledge about the application domain, or results of previous study, as a starting point for estimation and inference
- uninformative priors: avoid using external knowledge, "let the data speak for itself"

Weakly informative priors

A compromise between the informative and uninformative priors is so-called "weakly informative" priors, which generally attempt to include enough outside knowledge to ensure that the prior is proper and produces sensible results, but the information in the prior is often intentionally weaker than the available outside information.

- Our basketball prior example: I asked you to come up with rough bounds, but we used them loosely
- in the coin flip problem, take $\mathrm{Beta}(3, 3)$ in place of uniform or $\mathrm{Beta}(1, 1)$.
- Pfizer's $\mathrm{Beta}(0.7, 1)$ prior: compatible with vaccine efficacy between about -26 and 0.995, prior mean 0.3

Reference for prior choice: Stan Users Wiki (linked on D2L)

Summary

Today:

- Normal model, unknown variance
- More on priors
- Sampling and predictive checks

Next week:

- Short week (no class Monday)
- Regression models
- First look at PyMC3?