Multiparameter models and more on priors

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

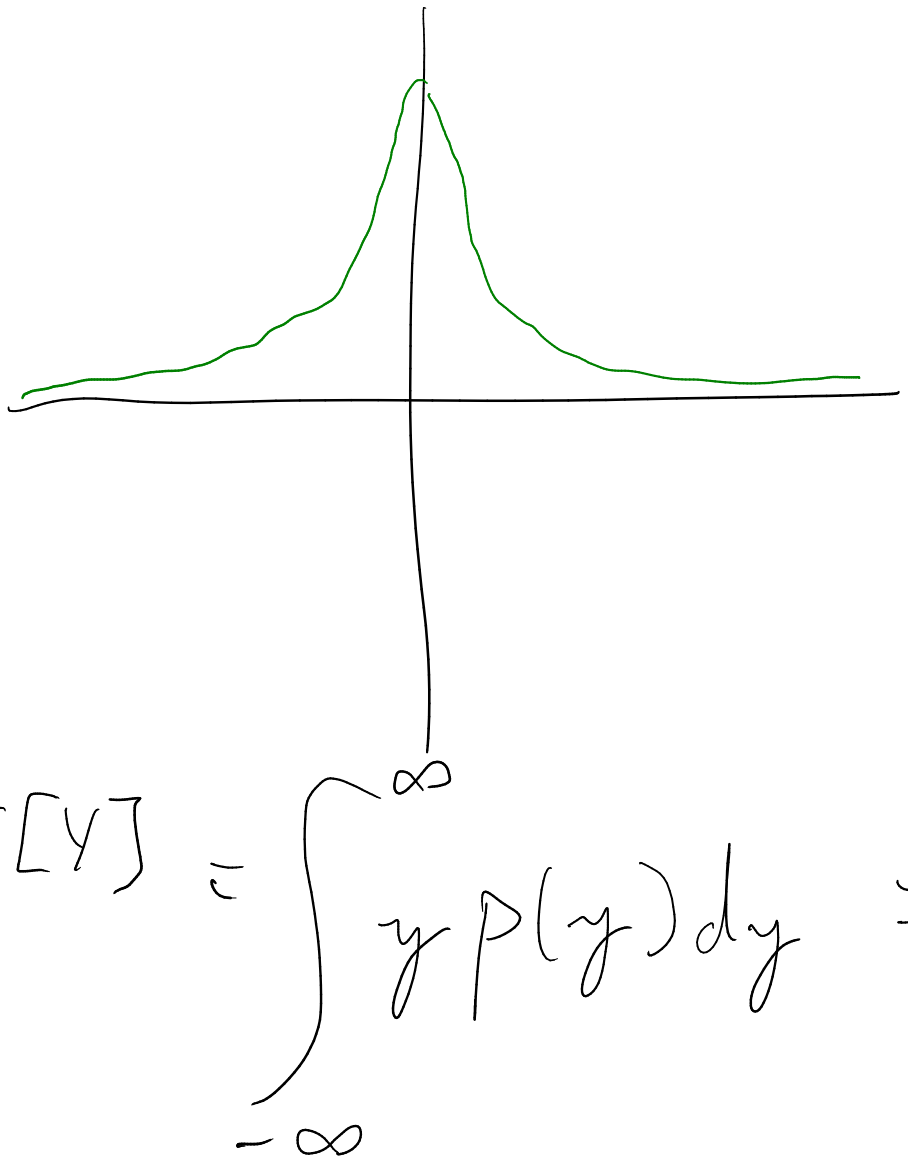
University of Arizona School of Information

February 10, 2021

Cauchy: $p(y \mid \theta, \gamma)$

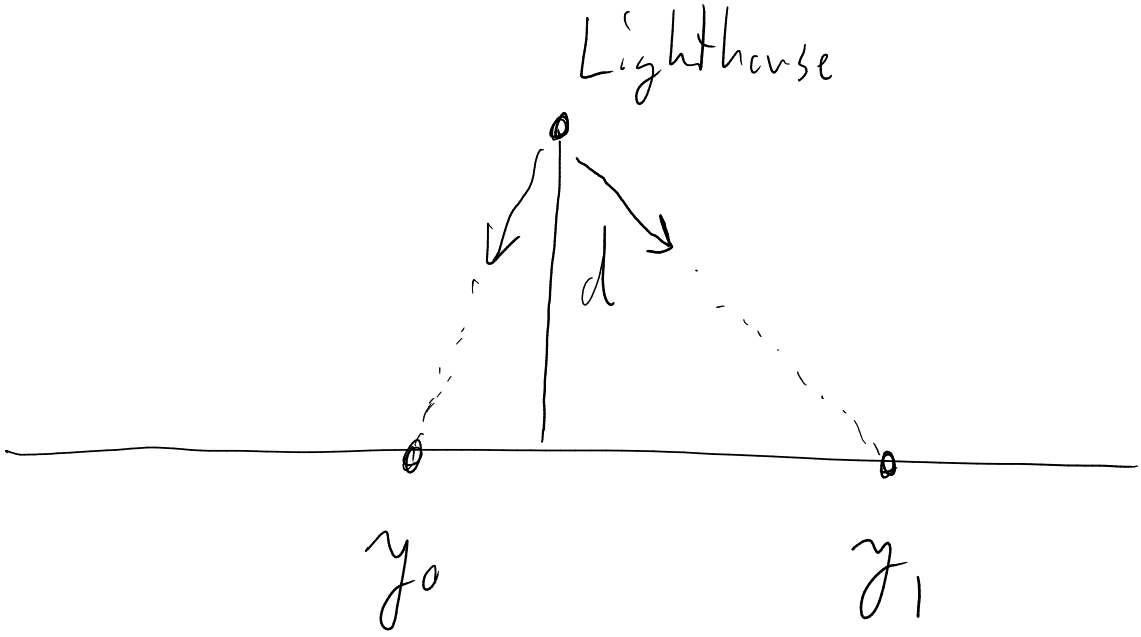
$$= \frac{1}{\pi\gamma\left(1 + \frac{y-\theta}{\gamma}\right)^2}$$

$\theta = 0, \gamma = 1$ $\quad p(y) = \frac{1}{\pi(1+y)^2}$



$$E[Y] = \int_{-\infty}^{\infty} y\,p(y)\,dy = \infty$$

Example | Lighthouse problem



$$p(\theta \mid y_1, y_2, \dots y_5)$$

$$= p(y_1 \mid \theta)\, p(y_2 \mid \theta) \dots p(y_5 \mid \theta)$$

$$= \frac{1}{\pi\left(1+(y_1-\theta)^2\right)} \cdot \frac{1}{\pi\left(1+(y_2-\theta)^2\right)} \cdot \dots \frac{1}{\pi\left(1+(y_5-\theta)^2\right)}$$

restricted to $[0, 100]$ by $p(\theta)$

sp.stats.cauchy(loc = θ, scale = 1)

Outline

Last time:

- Approximating a posterior
- Random sampling
- Normal model, unknown variance (a little)

Today:

- A little more on the normal model
- More on priors for multiparameter models

Normal model, unknown variance

Introduction to multi-parameter models

The known-variance assumption isn't necessarily particularly realistic. So instead, we can allow σ^2 to be an unknown parameter in our model.

New model (simple, improper priors):

$$y_i \sim \text{Normal}(\mu, \sigma^2)$$
$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

← noninformative prior

This improper prior is derived from applying a uniform prior to $\mu, \log \sigma$.

joint prior $p(\mu)p(\sigma^2)$

Priors for location and scale parameters

§ 2.9

Motivation for uniform prior on μ:

- The likelihood $p(y|\mu)$ has the property that $p(y-\mu|\mu)$ is not (explicitly) a function of y or μ
- μ is called a pure *location parameter*

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

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

Priors for location and scale parameters

Motivation for uniform prior on μ:

- The likelihood $p(y|\mu)$ has the property that $p(y-\mu|\mu)$ is not (explicitly) a function of y or μ
- μ is called a pure *location parameter*
- If we're looking for a non-informative prior, this property should appear in the posterior as well: so, $p(y-\mu|y)$ is a function only of $y-\mu$
- Since $p(\mu|y) \propto p(\mu)p(y|\mu)$, this requires that $p(\mu)$ does not explicitly depend on μ

$$p(\mu) \propto 1.$$

Priors for location and scale parameters

$$z = \frac{y - \mu}{\sigma}$$

Motivation for prior on σ^2:

- Similar reasoning, but replace $y - \mu$ with $\frac{y}{\sigma}$
- σ is called a pure *scale parameter*
- In this case, we get $p(\sigma) \propto \sigma^{-1}$ or $p(\log \sigma) \propto 1$

Note: when we specify a prior with multiple parameters, we're specifying a joint distribution – though in many cases parameters may be *a priori* independent

See conjugate prior for this model, §3.4?

Priors for location and scale parameters

This generalizes a bit the parameterization of the normal distribution in terms of mean and standard deviation:

- Want parameters that behave like the normal mean/SD, but may not be mean/SD
- Example 1: t distribution
  - Can say that if $\frac{y-\theta}{s} \sim t_\nu$, y has a t distribution with location θ and scale s
  - but, t_ν doesn't have standard deviation 1
- Example 2: Cauchy distribution (cf. exercise 2.11) uses location/scale parameters, but does not even *have* a mean or SD

The joint posterior

As before we can get the joint posterior by simply multiplying this prior by the likelihood $N(\mu, \sigma^2)$ for our data $y_1, \ldots, y_n$.

from prior. from Normal likelihood

$$p(\mu, \sigma^2|y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

$$= \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)]\right)$$

where $\bar{y}$ is the sample mean, s^2 the sample variance.

Interpreting the posterior

The posterior here shows that μ, σ are not independent.

In order to interpret this, it's easier to think about the distribution of μ conditional on σ^2. This is simple:

$$\mu|\sigma^2 \sim \text{Normal}(\bar{y}, \sigma^2/n)$$

which is the same as the known variance case. To factor the joint posterior, then, we need $p(\sigma^2|y)$.

$p(\mu|\sigma^2)\, p(\sigma^2)$

Marginal distribution of σ^2

The marginal posterior distribution of σ^2 is obtained by averaging the joint posterior over μ:

$$p(\sigma^2|y) \propto \int_{-\infty}^{\infty} \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu$$

Looks gnarly, but it's not – the exponential factors into the part dependent on s and the part dependent on μ. The part dependent on μ is a Gaussian integral, proportional to σ^{-1}. So, we get...

$$p(\sigma^2|y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2} n(y-\mu)^2\right) d\mu$$

prop to $\frac{1}{\sigma}$

Marginal distribution of σ^2

The marginal posterior distribution of σ^2 is

$$p(\sigma^2|y) \propto \sigma^{n-3} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)$$

which is a standard density on σ^2:

$$\sigma^2|y \sim \text{Inv}-\chi^2(n-1, s^2)$$

scaled inverse χ^2 distribution

Sampling from the joint posterior

It is easy to sample from the joint posterior distribution:

- draw a value of σ^2 from the posterior for σ^2 (a scaled inverse chi-squared distribution)
- draw a value of μ from the conditional posterior given your value of σ^2 (a normal distribution)

Sampling from the joint posterior

Let's draw and interpret some samples from the posterior for the basketball score variable.

Now, we take y_i to be the total score in a game, and specify the model:

$$y_i \sim \text{Normal}(\mu, \sigma^2)$$
$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

We'll start by plotting points in the (μ, σ) plane, then the normal distributions they represent

Sampling from the joint posterior

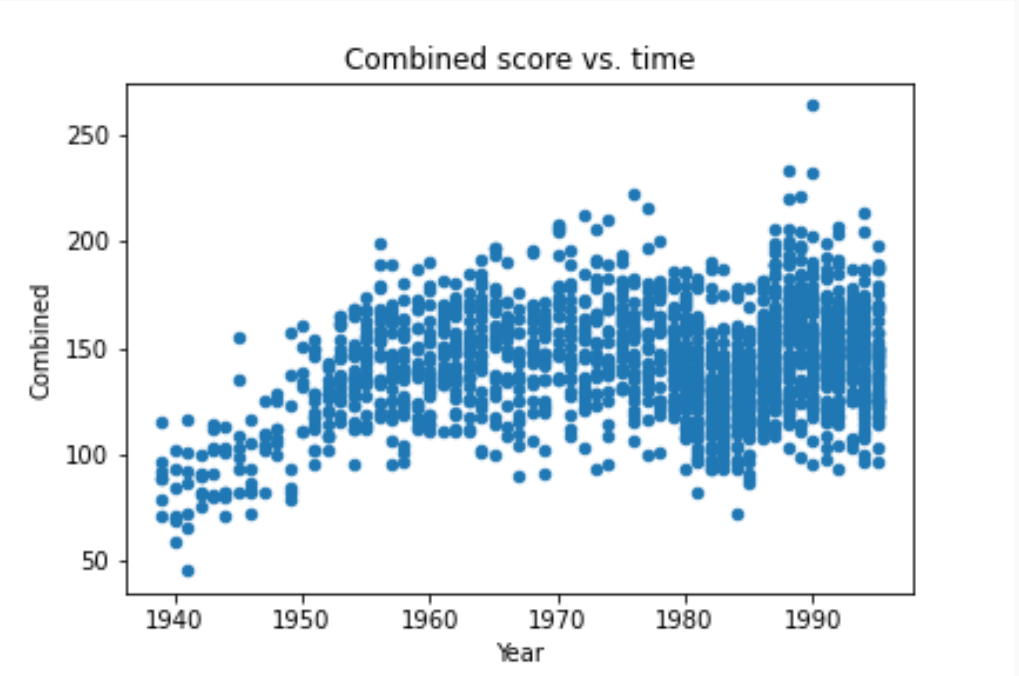
For the following plots:

- Restricted to year ≥ 1988 for consistency in the target distribution
- Subsampled: $n = 10, 40, 100$ to see how sample size affects uncertainty

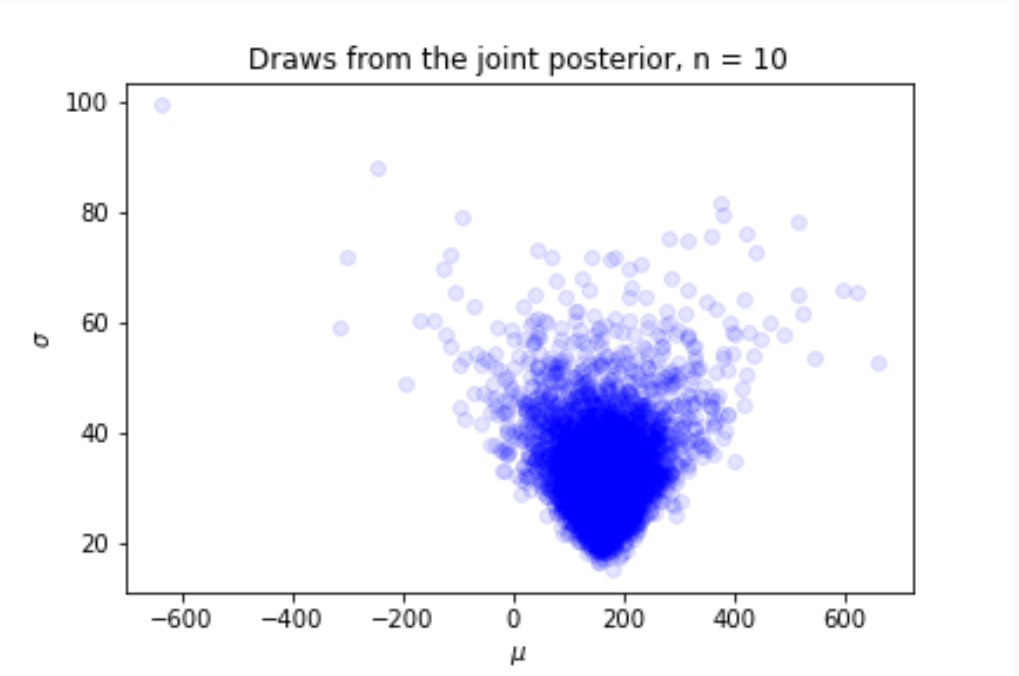
Why limit to data after 1988?

Time series

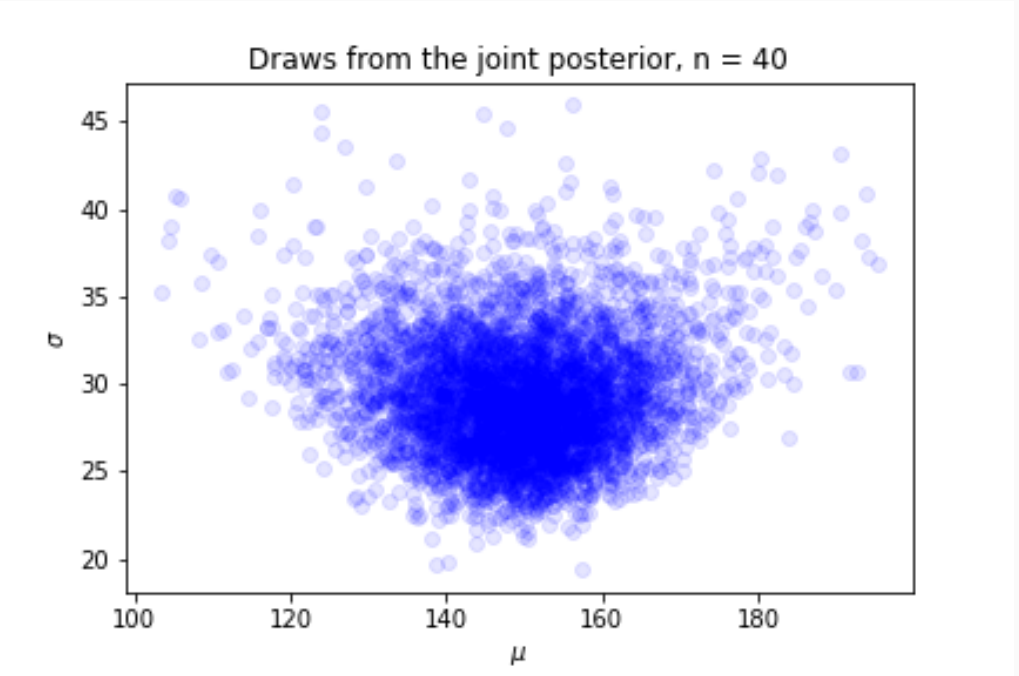
Plotting all of the combined scores against time:



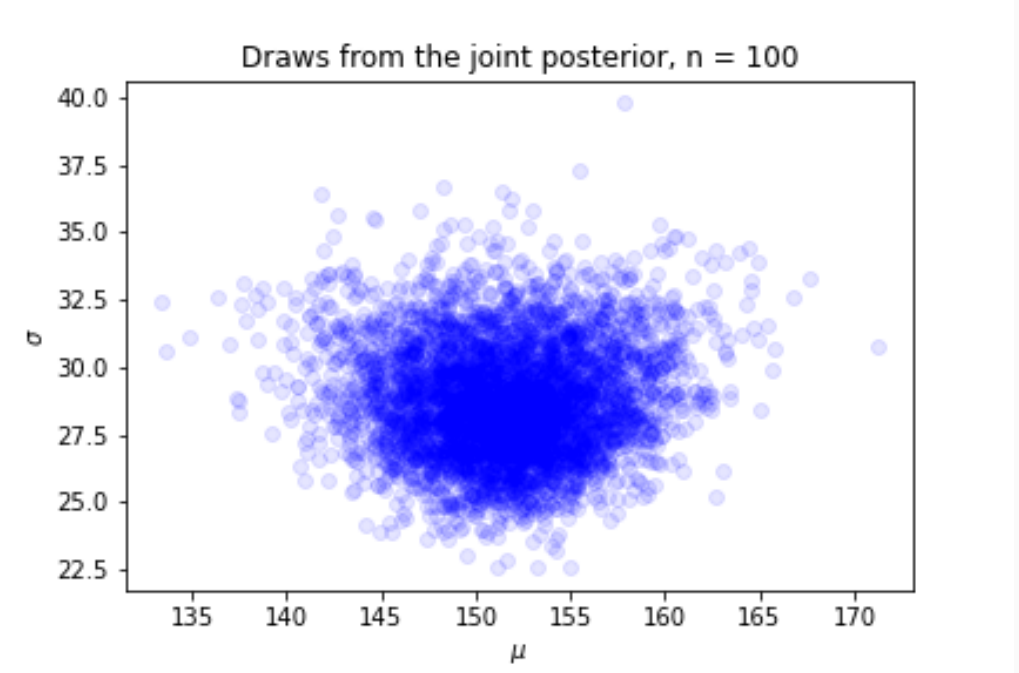
Sampling from the joint posterior



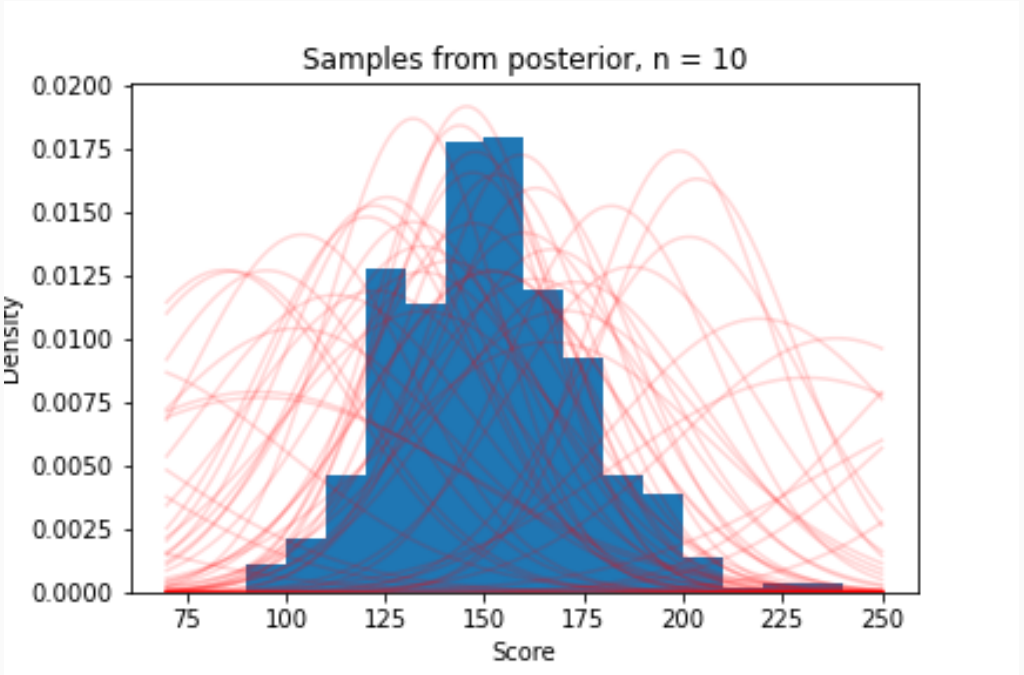
Sampling from the joint posterior



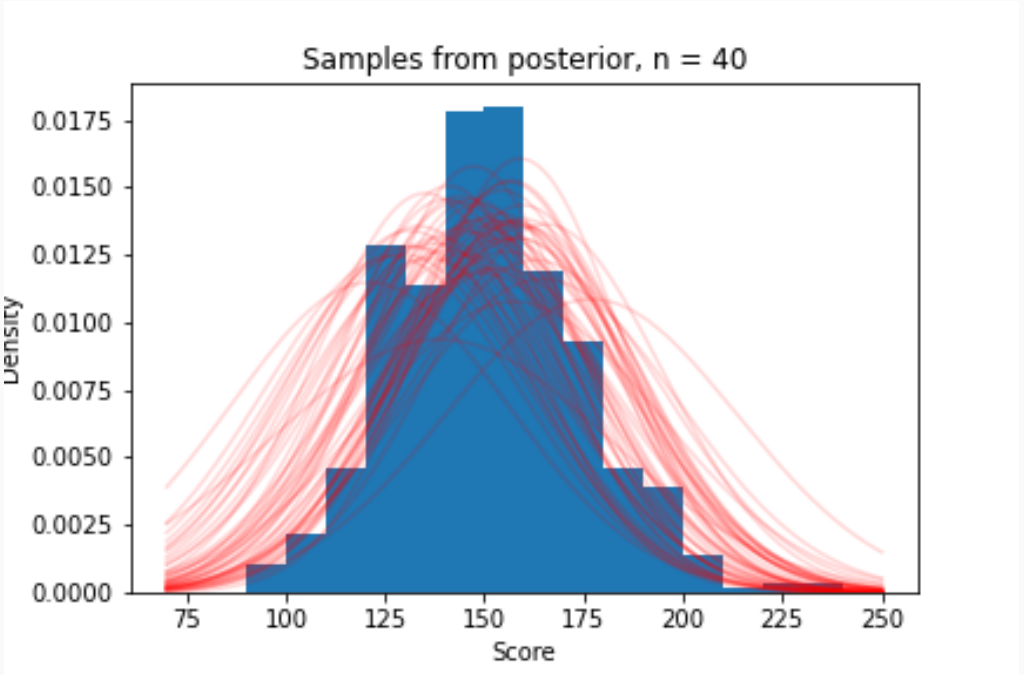
Sampling from the joint posterior



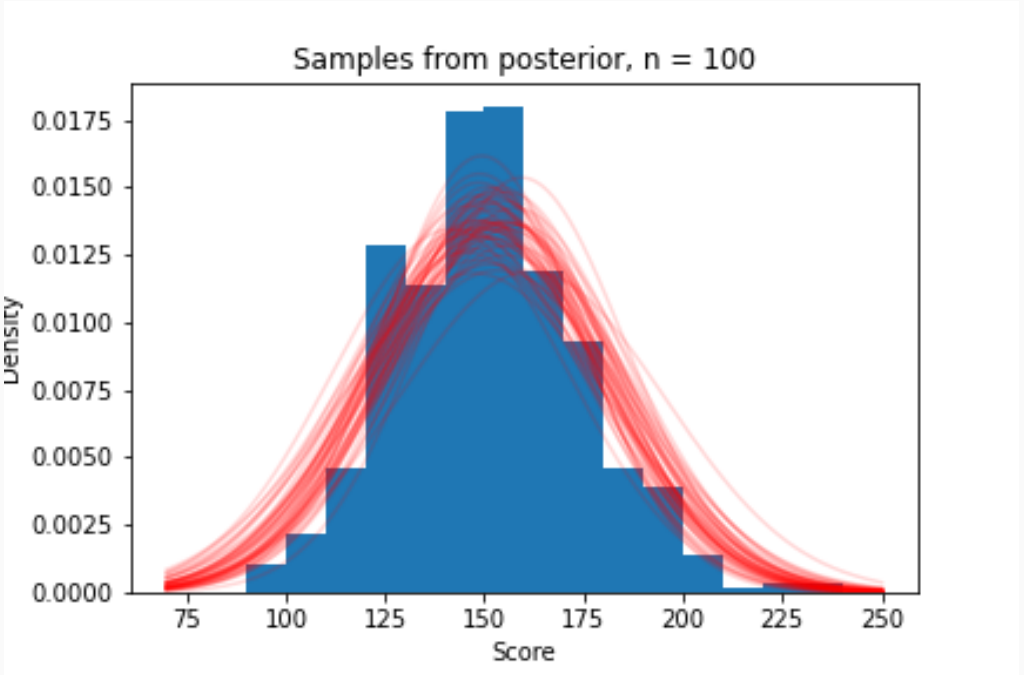
Visualizing uncertainty



Visualizing uncertainty



Visualizing uncertainty



Another example: speed-of-light data

Newcomb's speed of light data

1878: Simon Newcomb plans to measure the speed of light

- Measure the time for light to travel from the US Naval Observatory to a mirror at the Washington Monument – about 7442 m
- Travel times recorded as deviations from 24,800 ns

Let's apply the normal model:

$$y_i \sim \text{Normal}(\mu, \sigma^2)$$
$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

More results on the normal model

We'll do the same computations, but first:

- Can we get analytical formulas for distributions of interest?
- Can we assess whether the normal model was a good choice here?

Formulas

Previously we wrote down the joint posterior for σ^2, μ; in many cases, what we're most interested in is μ (e.g. when measuring the speed of light).

Derivation from pp.64-66 of the book leads to:

$$\mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

or,

$$\left.\frac{\mu - \bar{y}}{s/\sqrt{n}}\right| y \sim t_{n-1}$$

Trick is integration by substitution (change of variables) to eliminate the "nuisance" parameter σ^2

Assessing the normal model

There are a variety of qualitative and quantitative approaches to evaluating a model, but we'll start with a simple idea:

- Our model is generative – i.e., it gives a prescription for generating data given parameters
- We have estimates in the form of probability distributions for the parameters
- If the model fits the data, then it should be able to generate "replications" that reproduce properties of the real data

So, we sample from the posterior predictive distribution and compare it to the real data

Summary

Today:

- Two examples of the normal model with unknown variance
- Introduced the idea of posterior predictive checking

Next time:

- More posterior (and prior) predictive checks