Bayesian Modeling
Monte Carlo approximation

Yixin Wang

4 Monte Carlo approximation

Suppose that we are interested in quantities of interest for the posterior distribution, such as

(i) $\Pr(\theta \in A | y_1, \ldots, y_n)$ for some subset $A \subset \Theta$.

(ii) Posterior mean, variance, confidence intervals for $\theta_1 - \theta_2$, θ/θ_2, $\max\{\theta_1, \ldots, \theta_m\}$.

Under conjugacy, some of these quantities may be explicitly available in closed form, but this is not always the case. When we deal with complex models where no conjugate form of the prior is available, then posterior computation becomes a huge issue. This is in fact the main barrier for Bayesian statistics before the age of computers. Thankfully with the computational advances, such barrier can be overcome. One of the primary computational techniques for Bayesian computation is Markov Chain Monte Carlo. In this section, we will explore the "Monte Carlo" part of the technique.

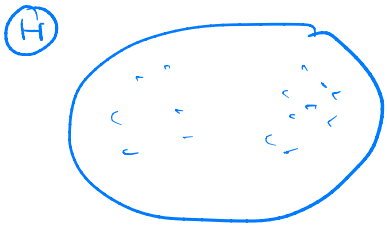
4.1 Basic ideas

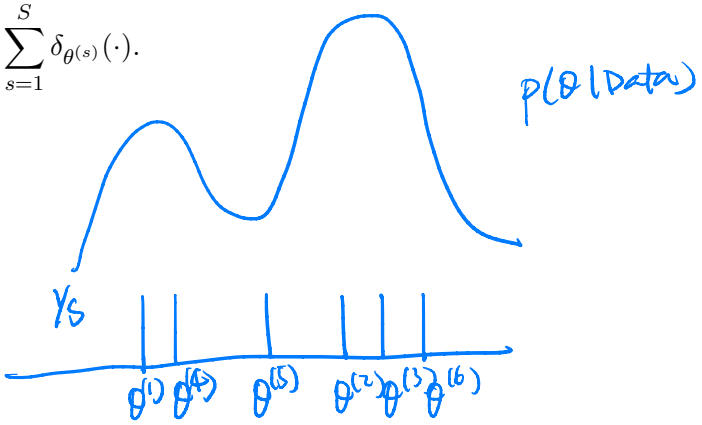
Suppose we could sample some number S of i.i.d. samples of the posterior distribution

$$\theta^{(1)}, \ldots, \theta^{(S)} \overset{iid}{\sim} p(\theta | y_1, \ldots, y_n).$$

Then the posterior distribution can be approximated with the empirical distribution provided by the S-sample. Notationally:

$$p(\cdot | y_1, \ldots, y_n) \approx \frac{1}{S} \sum_{s=1}^{S} \delta_{\theta^{(s)}}(\cdot).$$





To write down empirical dist. we use "Delta distribution" δ_θ

If a sample $\eta \sim \delta_\theta(\cdot)$

it means $\eta = \theta$

$\Rightarrow$ empirical distribution $\sum_{s=1}^{S} \frac{1}{S} \delta_{\theta^{(s)}}(\cdot)$

S: Monte Carlo sample size $\neq$ Data sample size n.

The Monte Carlo technique is simply this: take any function $g(\theta)$ (that is integrable with respect to the posterior distribution), by the law of large numbers, as $S \to \infty$,

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \to \mathbb{E}[g(\theta)|y_1, \ldots, y_n] = \int g(\theta)p(\theta|y_1, \ldots, y_n)\mathrm{d}\theta.$$

Take different choices for function g, we obtain

- $\bar{\theta} := \sum_{s=1}^{S} \theta^{(s)}/S \to \mathbb{E}[\theta|y_1, \ldots, y_n]$.
- $\frac{1}{S-1}\sum_{s=1}^{S}(\theta^{(s)} - \bar{\theta})^2 \to \mathrm{Var}[\theta|y_1, \ldots, y_n]$.
- $\#(\theta^{(s)} \le c)/S = \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(\theta^{(s)} \le c) \to \Pr(\theta \le c|y_1, \ldots, y_n)$.
- median$\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{1/2}$.
- the α-percentile of $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ tends to θ_α.

Numerical evaluation In the previous section, we use a Poisson sampling model, $Y_1, \ldots, Y_n | \theta \sim \text{Poisson}(\theta)$, and endow parameter θ with a gamma prior: $\gamma \sim \text{Gamma}(a, b)$. We know that the posterior of θ is $\text{Gamma}(a + \sum y_i, b + n)$, which yields the posterior mean $(a + \sum y_i)/(b + n) = 68/45 = 1.51$.

If we didn't have this mean formula, we can appeal to Monte Carlo approximation in R.

First, to obtain random Gamma samples

```
a<-2    ; b<-1
sy<-66  ; n<-44

theta.mc10<-rgamma(10,a+sy,b+n)
theta.mc100<-rgamma(100,a+sy,b+n)
theta.mc1000<-rgamma(1000,a+sy,b+n)
```

S=10
S=100
S=1000

To obtain the mean

```
> mean(theta.mc10)
[1] 1.532794
> mean(theta.mc100)
[1] 1.513947
> mean(theta.mc1000)
[1] 1.501015
```

← apply mean (expectation) to the S MC samples

and probabilities of intervals of interest

```
> mean(theta.mc10<1.75)
[1] 0.9
> mean(theta.mc100<1.75)
[1] 0.94
> mean(theta.mc1000<1.75)
[1] 0.899
```

S=10 ← $\Pr(\theta < 1.75 | y_1, \ldots, y_n)$
S=100
S=1000

or relevant quantiles

```
> quantile(theta.mc10, c(.025,.975))
    2.5%    97.5%
1.260291 1.750068
> quantile(theta.mc100, c(.025,.975))
    2.5%    97.5%
1.231646 1.813752
> quantile(theta.mc1000, c(.025,.975))
    2.5%    97.5%
1.180194 1.892473
```

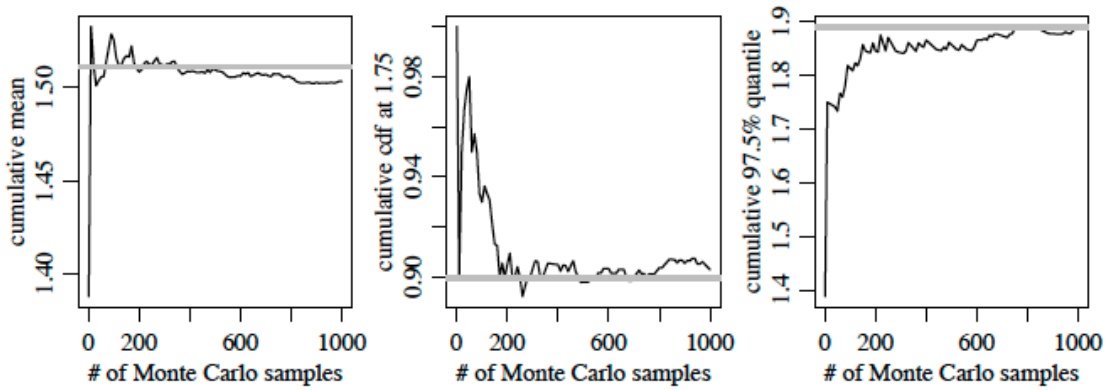


Figure 4.1: Convergence of Monte Carlo estimates as MC sample size increases.

Fig. 4.1 provides an illustration of the effects of increasing Monte Carlo sample size S. Note that the MC sample size S has nothing to do with the sample size of the data set given/observed. S represents the computational cost, which becomes cheaper as the computer becomes more powerful.

To standard way of choosing S is to choose it just large enough so the Monte Carlo standard error is less than the precision to which we want to report the quantity of interest.

$\bar{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}$

Example 4.1. We want to compute the posterior expectation of θ. The Monte Carlo estimate gives us $\bar{\theta}$. By the central limit theorem, the sample mean $\bar{\theta}$ is approximately distributed as normal distribution with expectation $\mathbb{E}[\theta|y_1, \ldots, y_n]$ and standard deviation $\sqrt{\mathrm{Var}[\theta|y_1, \ldots, y_n]/S}$.

95%

-1.96 1.96

So letting $\hat{\sigma}^2 = \frac{1}{S-1}\sum(\theta^{(s)} - \bar{\theta})^2$ be the *MC estimate of the variance* σ^2, then MC standard error (of the MC estimate of the posterior mean for θ) is $\sqrt{\hat{\sigma}^2/S}$. Thus, the approx. 95% MC confidence interval for the posterior mean is $\bar{\theta} \pm 2\sqrt{\hat{\sigma}^2/S}$.

For example, one set $S = 100$ and found that the MC estimate of $\mathrm{Var}[\theta|y_1, \ldots, y_n]$ was 0.024. Then the approximate MC standard error for the mean would be $\sqrt{0.024/100} = 0.015$. Suppose that you wanted the difference between the posterior mean $\mathbb{E}[\theta|y_1, \ldots, y_n]$ and its MC estimate to be less than 0.01 with high probability (i.e., $> 95\%$ confidence) then you would need to increase your sample size so that $2\sqrt{0.024/S} < 0.01$, i.e., $S > 960$. □

MC error tolerance level

MC stderr

4.2 Posterior inference for arbitrary functions

Recall the example of birthrates in Section 3.4. Based on the prior specifications and the data of birthrates, the posterior distributions for the two educational groups are

$$\begin{aligned} \{\theta_1|y_{1,1},\ldots,y_{n_1,1}\} &\sim \text{Gamma}(219,112) \quad \text{(women without bachelor's degrees)} \\ \{\theta_2|y_{1,2},\ldots,y_{n_2,2}\} &\sim \text{Gamma}(68,45) \quad \text{(women with bachelor's degrees).} \end{aligned}$$

We are interested in $\Pr(\theta_1 > \theta_2|\text{Data from both groups})$, or the posterior of the ratio θ_1/θ_2. Obtain Monte Carlo samples *independently* for the two data groups:

$$\begin{aligned} \text{sample } \theta_1^{(1)},\ldots,\theta_1^{(S)} &\overset{iid}{\sim} p(\theta_1|\text{Data from first group}), \\ \text{sample } \theta_2^{(1)},\ldots,\theta_2^{(S)} &\overset{iid}{\sim} p(\theta_1|\text{Data from second group}). \end{aligned}$$

Accordingly, the pairs of $(\theta_1^{(s)},\theta_2^{(s)})$ for $s=1,\ldots,S$ are i.i.d. Monte Carlo samples. We can approximate

$$\Pr(\theta_1 > \theta_2|\text{Data from both groups}) :\approx \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(\theta_1^{(s)} > \theta_2^{(s)}).$$

In R codes

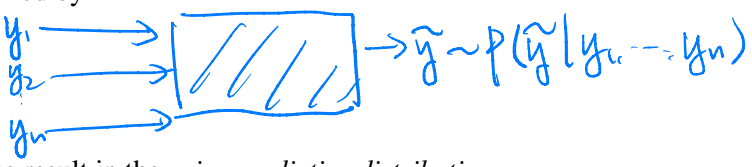
```
> a<-2 ; b<-1
> sy1<-217 ;  n1<-111
> sy2<-66  ;  n2<-44

> theta1.mc<-rgamma(10000,a+sy1, b+n1)
> theta2.mc<-rgamma(10000,a+sy2, b+n2)

> mean(theta1.mc>theta2.mc)

[1] 0.9708
```

4.3 Sampling from posterior predictive distributions

Parameter θ and the prior on θ represent the modeler's understanding of the data population. Different modelers may come up with different parameterization and different prior specification. How do we verify the validity and compare among different models? This is usually done through assessment of the predictive distribution.

We saw examples of predictive distributions in Section 3. In general, a predictive distribution is the (marginal) distribution of unobserved data $\tilde{Y}$ which is obtained by

- having all known quantities been conditioned on;
- having all unknown quantities been integrated out.



Before we have seen any data, all modeling assumptions result in the *prior predictive distribution*

$$p(\tilde{y}) = \int p(\tilde{y}|\theta)p(\theta)\mathrm{d}\theta.$$

Having observed the data set $\{y_1, \ldots, y_n\}$, we obtain the *posterior predictive distribution*

$$\begin{aligned} p(\tilde{y}|y_1, \ldots, y_n) &= \int p(\tilde{y}|\theta, y_1, \ldots, y_n)p(\theta|y_1, \ldots, y_n)\mathrm{d}\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y_1, \ldots, y_n)\mathrm{d}\theta. \end{aligned}$$

Example 4.2. Continue on the birth rates modeling considered earlier. We assumed a Poisson sampling model: $Y|\theta \sim \text{Poisson}(\theta)$ for a data population (say the group of women aged 40 with a college degree). We placed a Gamma prior on θ: $\theta \sim \text{Gamma}(a, b)$. We found that the resulting prior predictive distribution of $\tilde{Y}$ is a negative binomial (a, b).

Having observed an n-data sample, we found that the posterior distribution of θ is $\text{Gamma}(a+\sum y_i, b+n)$, and the predictive distribution of $\tilde{Y}$ is a negative binomial with parameters $(a + \sum y_i, b + n)$. In this example, thanks to conjugacy we have a very closed form for the predictive distribution, both a priori and a posteriori.

In general, we probably won't be so "lucky" — most realistic models do not admit a closed form for the posterior distributions. In order to evaluate the posterior predictive distributions, we may proceed by drawing samples from them instead.

$$p(\tilde{y}|y_1,\cdots,y_n) = \int p(\tilde{y}|\theta)\cdot p(\theta|y_1,\cdots,y_n)\,d\theta$$

poisson Gamma

giving the weights to the mixture.

The key is to observe that $p(\tilde{y}|y_1,\ldots,y_n)$ can be viewed as a *mixture* of the sampling distributions $p(\tilde{y}|\theta)$, where the θ is randomly mixed by the posterior distribution $p(\theta|y_1,\ldots,y_n)$. If we can draw samples from the posterior of θ, we can use such samples to again draw samples from the sampling distribution, with each θ given.

To be specific, for $s = 1,\ldots,S$ obtain independent Monte Carlo samples as follows

- sample $\theta^{(s)} \sim p(\theta|y_1,\ldots,y_n)$, and then sample $\tilde{y}^{(s)} \sim p(\tilde{y}|\theta^{(s)})$.

Then, we have obtained a valid i.i.d. n-sample $\tilde{y}^{(1)},\ldots,\tilde{y}^{(S)}$ from the posterior predictive distribution.

Example 4.3. Continue on the birth rates modeling example. Suppose we are interested in the predictive probability that an age-40 woman without a college degree wold have more children than an age-40 woman with a college degree (using prior Gamma parameters $a = 2, b = 1$):

posterior predictive

NegBinom $(a+\sum y_i,\ b+n)$

$$\Pr(\tilde{Y}_1 > \tilde{Y}_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66)$$

$$= \sum_{\tilde{y}_2=0}^{\infty}\sum_{\tilde{y}_1=\tilde{y}_2+1}^{\infty} \text{NegBinomial}(\tilde{y}_1, 219, 112)\times \text{NegBinomial}(\tilde{y}_2, 68, 45).$$

$P(\tilde{Y}_1=\tilde{y}_1,\ \tilde{Y}_2=\tilde{y}_2 \mid \sum Y_{i1}=217,\ \sum Y_{i2}=66)$

This can be easily evaluated via the MC technique. In R codes

```
> a<-2 ; b<-1
> sy1<-217 ;   n1<-111
> sy2<-66  ;   n2<-44

> theta1.mc<-rgamma(10000,a+sy1 , b+n1)
> theta2.mc<-rgamma(10000,a+sy2 , b+n2)
> y1.mc<-rpois(10000,theta1.mc)
> y2.mc<-rpois(10000,theta2.mc)

> mean(y1.mc>y2.mc)
[1] 0.4823
```

$S = 10000$

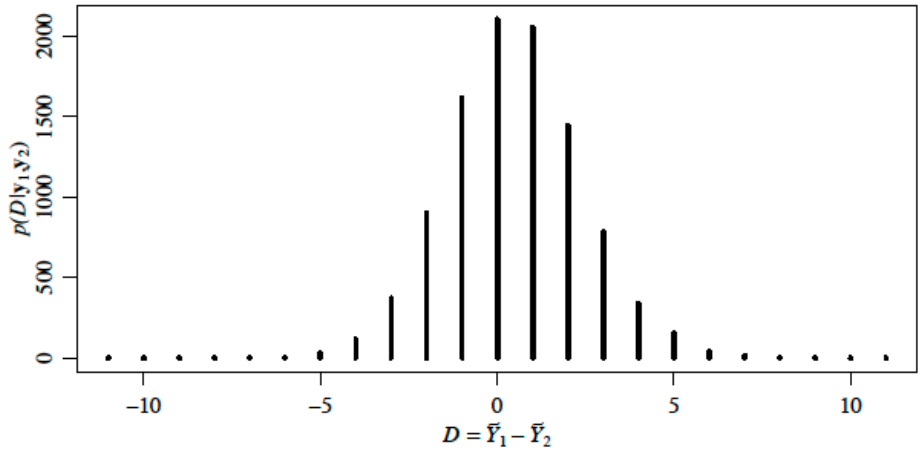
$(\theta_1^1,\cdots,\theta_1^S) \sim p(\theta \mid \text{Data group 1})$

$(\theta_2^1,\cdots,\theta_2^S) \sim p(\theta \mid \text{Data group 2})$

$(\tilde{Y}_1^1,\cdots,\tilde{Y}_1^S) \sim$ Mixture of Poisson samples arising from group 1

$(\tilde{Y}_2^1,\cdots,\tilde{Y}_2^S) \sim$ Mixture of Poisson samples from group 2.

We can also compute other quantities of interest based on these MC samples. We can also plot an estimate of the posterior predictive distribution for $\tilde{Y}_1 - \tilde{Y}_2$, as illustrated in Fig. **??**.



Additional remark We can use the same technique to draw samples for prior predictive distribution; such samples are then utilized for setting prior parameters. This technique is very useful if the prior distribution is not conjugate, and/or the prior predictive distribution is not easily accessible via closed form expressions.

prior predictive distribution is not too far from the data

loose sanity check.

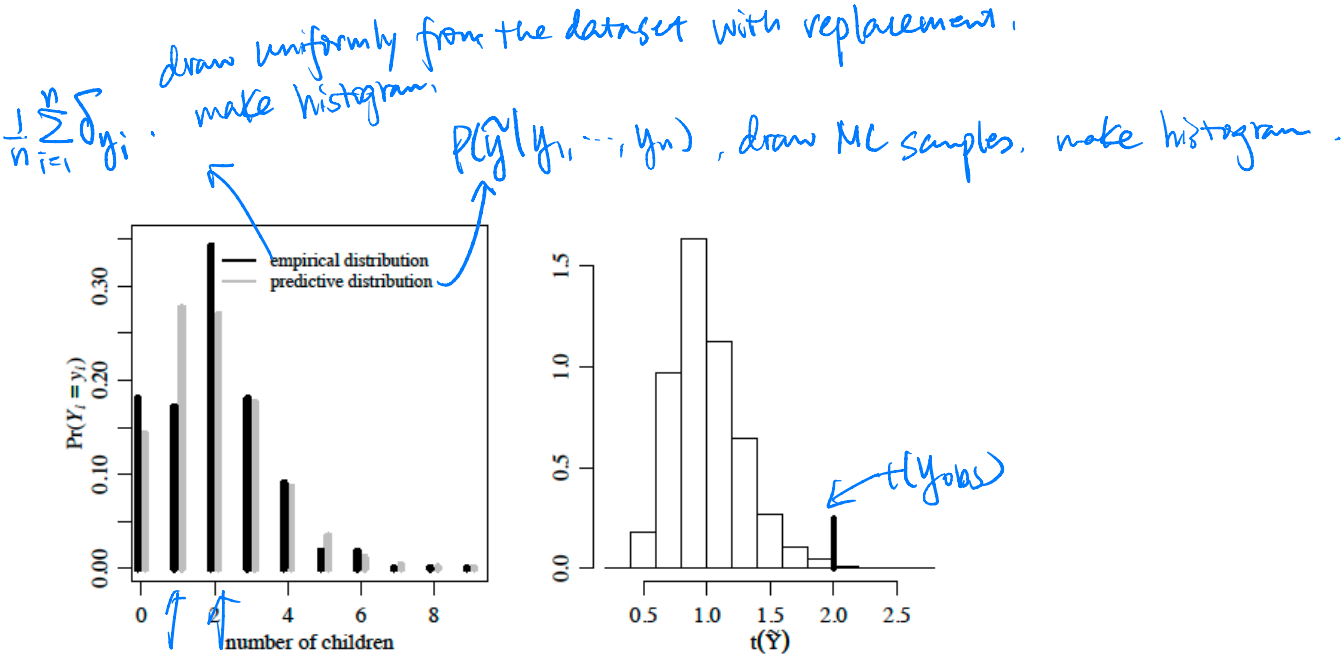


Figure 4.2: Evaluation of model fit. Left panel: the empirical and predictive distributions of the number of children of women without a bachelor's degree. Right panel: The posterior predictive distribution of the empirical odds of having two children versus one child in a data set of size $n_1 = 111$. The observed odds are given in the short vertical line.

4.4 Posterior predictive model checking

We again use the birthrates data example to illustrate the important issue of model checking via posterior predictive distributions. We used a Poisson sampling model endowed with a Gamma prior to describe the number of children of groups of age-40 women with or without college degrees.

Consider the group of women without college degrees, for which we arrived at the posterior predictive distribution for $\tilde{Y}_1$ (which is a negative binomial). Let us compare that distribution with the empirical distribution. Note that these are two products that are computed out of the same data sample $\{y_{1,1}, \ldots, y_{n_1,1}\}$, where $n_1 = 111$.

In the empirical sample, shown in back, the number of women with exactly two children is 38, which is twice the number of women with one child. By contrast, this group's posterior predictive distribution, shown in gray, suggests that the probability of sampling a woman with two children is slightly less than of sampling a woman with one (0.27 and 0.28, respectively). How do we make sense of this significant discrepancy?

There are two possible explanations.

- There is a sampling variability and the sample size is probably too small, so the empirical distribution of sampled data does not generally match exactly the distribution of the population. In fact, empirical distributions (like all histograms) usually look bumpy, so having a predictive distribution that smoothes over the bumps may be desirable.
- An alternative explanation is that the Poisson model is quite wrong. This is plausible because there is no Poisson distribution with such a sharp peak at $y = 2$. Having said that, note that the posterior predictive distribution is in fact a mixture of Poissons that equals a negative binomial, so this explanation needs further evaluation.

"test statistic"

which aspect of posterior predictive distribution we want to focus the check on

We can evaluate the validity of the posterior predictive model via Monte Carlo simulation. We need a "marker", and in this case we use the ratio of the number of $y = 2$'s to the number of $y = 1$'s in our data. For every vector $\boldsymbol{y}$ of length $n_1 = 111$, let $t(\boldsymbol{y})$ denote this ratio. For our observed data sample, $\boldsymbol{y}_{\text{obs}}$, we have $t(\boldsymbol{y}_{\text{obs}}) = 2$.

↳ vector of all data points

What sort of values of $t(\tilde{\boldsymbol{Y}})$ should one expect, if $\tilde{\boldsymbol{Y}}$ are drawn from the posterior predictive distribution? The Monte Carlo simulation procedure is as follows. For $s = 1, \ldots, S$,

- sample $\theta^{(s)} \sim p(\theta|\boldsymbol{Y} = \boldsymbol{y}_{\text{obs}})$.
- sample $\tilde{Y}^{(s)} = (\tilde{y}_1^{(s)}, \ldots, \tilde{y}_{n_1}^{(s)}) \overset{iid}{\sim} p(y|\theta^{(s)})$.
- compute $t^{(s)} = t(\tilde{Y}^{(s)})$.

↳ drawn a MC sample with same size as the dataset from posterior predictive

The right panel of Fig. 4.2 shows the histogram of $t(\tilde{Y})$ that one can get out of 10000 Monte Carlo samples (note: each MC sample here consists of an n_1-sample represented by $\tilde{Y}^{(s)}$). Observe that out of 10000 such datasets only about 0.5% had values of $t(\boldsymbol{y})$ that equaled or exceeded $t(\boldsymbol{y}_{\text{obs}})$. This indicates that our Poisson sampling model is flawed. If one is in particular interested in a more accurate model for Y, perhaps a complex sampling model than the Poisson is warranted.

same size as the original dataset

Certain aspects of the Poisson sampling model that may still be useful in this example. For instance, if we are only interested in population parameters such as the mean and variance via θ, then Poisson is quite accurate in capturing the relationship between these quantities, as the empirical mean and empirical variance is found to be 1.95 and 1.90, respectively.

It is known in theory that even if a model is misspecified, some aspects of the population may still be accurately estimated with such a model. In practice, as George Box said, all models are wrong, but some are useful. Thus, while statistical modelers constantly search for better models, and we have a vast arsenal for doing so as you will see in later lectures, we do not readily discard simpler ones just for the sake of bigger and more complex models.

The materials in this course are adapted from materials created by David Blei, Yang Chen, Andrew Gelman, Scott Linderman, Long Nguyen, and the 3blue1brown channel.