Bayesian Evidence Synthesis:opioid crisis

Hyeongcheol Park* & Paul Gustafson* & Micheal A Irvine*[1]

February 2, 2020

Contents

List of Figures

List of Tables

Abstract

1 Introduction

The opioid crisis is a major issue in North America including British Columbia, Canada. British Columbia is in the midst of a drug overdose crisis due to illicit opioids and the circumstance s ~~is~~ are getting worse; . According to the BC Coroners Service, there were more than 930 apparent illicit drug overdose deaths in BC from Jan. 1 to Dec. 31, 2016. This compares to 518 in 2015, an increase of 79.2%. [1] The goal of this project is to apply Bayesian evidence synthesis to understand better the opoid crisis in British Columbia, Canada.

One difficulty to cope in coping with the opioid crisis is that the total number of overdoses ~~is~~ unknown. The reason is that the use of prescription opioids often leads to use of illicit opioids and the illicit usage cannot be known. Many of those who become addicted to opioids do so after initially receiving a prescription. The highly addictive nature of these pain relievers makes it easy for the human brain to crave more. It is only after their prescription ends that many users realize they've become dependent on the effects of opioids to function "normally." At that point, they are either forced to get clean and endure the pain that comes with the withdrawal symptoms of opioids or look for another means of getting their high. This is often the time where people will turn to illicit drugs or other analogues. Because prescription opioids are so expensive, this is when many users turn to heroin. It is often cheaper, more potent, and easier to locate than what they were taking before. In fact, about 80% of people using heroin started with a prescription to another opioid. After using heroin, however, 23% of individuals develop opioid addiction.[2] [attempt: need to give a backstory-why isn't this observable? The number of overdoses]

Since the total number of overdoses ~~are~~ is unknown, ~~the~~ this number needs to be estimated. Bayesian statistics can be a way to approach the problem and give us the a good estimate of the number. Bayesian statistics is a theory in the field of statistics based on the Bayesian interpretation of probability where probability expresses a degree of belief in an event. [3] Detailed explanation is provided at the following section. [attempt: explain Bayesian statistics]

defensible? (depending on the data, a precise estimate may not be possible)

All examples here were performed in Python 3.7 using the library pyMC (reference) and JAGS (reference). Training was performed using No U-Turn Sampling (NUTS) over two chains with 1000 iterations (is it sample size?). Fitting was performed on a GHz Intel Core i5 with 8GM of LPDD3 RAM and typically had wall times under ten minutes. Data processing was carried out using the Pandas and SciPy library [reference]. Data visualization was performed using the libraries Seaborn and Matplotlib [ref]. Code for all examples in this study are provided.

2 methods

M

target?

The number of overdoses is our ultimate interest of estimation. To achive the goal in context of Bayesian statistics, we use our prior belief and available data set about the target of estimation. The posterior belief is coming from both parts of the information. Following equation express the idea in a mathmatical form:

these sources of ?

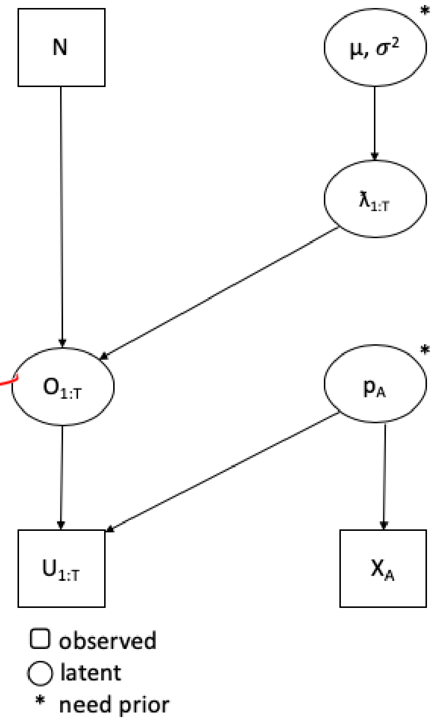
of The

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \propto p(\theta)p(y|\theta) \tag{1}$$

θ = all latent variables and params?

where θ represents the target of estimation, the total overdoses and y represents the collected data set providing relevant samples of the target variable. In Bayesian concepts, the variable of interest θ is a fixed number but the number is yielded from a certain distribution; That is, θ is considered to be a sample of a random variable. The problem, however, is that the data y is not available. We can have our prior belief about the total overdoses but it is not possible to obtain any data. Hence it is needed to approach the estimation in a indirect fashion. Figure 1 shows an example of how the estimation can be proceeded. [attempt: Give a graphical model representation? state that(?)]

θ fixed but distribution describes knowledge about θ



have to define all or none of the nodes in the caption (none would be simpler)

Figure 1: Example of estimating the total overdoses indirectly given two data sets: ambulance attended overdoses and survey data. N is the population size

expand more - witnesses to overdoses being surveyed? confirming when someone called ambulance

Let O_t the number of overdoses in a given month t in British Columbia. [attempt: specify a geographic area] Suppose there was a survey conducted to estimate the proportion of ambulance call p_A among the overdoses. p_A is assumed constant across time for simplicity. Let n_A the sample size of the survey and x_A to be the total number who confirmed they did call ambulance. It is assumed that x_A follows a Binomial distribution:

$$x_A \sim Bin(n_A, p_A) \text{ ambulance call-outs model} \tag{2}$$

more natural to switch order?

The total overdoses need to be modeled. The simplest conceptual model is to take an underlying log-rate z_t that is independent and identically distributed across months according to a normal distribution with mean μ and variance σ^2. [4] Denote λ_t^{OD} the rate of overdose at time t. It is assumed that the total overdose O_t follows Poisson distribution where the population of the region of interest is N.

$$\left.\begin{aligned} z_t &\sim N(\mu, \sigma^2) \\ \lambda_t^{OD} &= \exp(z_t) \\ O_t &\sim Poi(\lambda_t^{OD} N) \end{aligned}\right\} \text{ overdose model} \tag{3}$$

Estimation of O_t is not straightforward since none of the variables (μ, σ, N) determining O_t is known. Hence O_t should be inferred from using U_t and p_A, where p_A is the ambulance call out rate and U_t is the number of ambulance-attended overdoses at a time point t. In general, the data of ambulance-attended overdoses U_t can be obtained. It is assumed that U_t follows Binomial distribution:

initially at least treating N as known

$$U_t \sim Bin(O_t, p_A) \tag{4}$$

Now O_t can be estimated as p_A can be infered by survey data and the data regarding U_t is given. We suggest a simple model as a start where the model only combines Ambulance Call-outs Model (2) and Overdose Model (3).

The next step is to run some simulations to figure out how different types of inputs lead some changes of output. To do so, the simple model is illustrated below.

in Fig. 1?

2.1 Simulation

The first simulation simplifies the assumptions of variables as much as possible; we assumed $N = 10000, n_A = 1000$. Note that the population size, N, is unknown in reality but it is assumed known for the investigator here for simplicity as in figure 1. The assumptions will change later to see the impact of the likelihood over the posterior distributions of variables of interest; the total population size for a region, N, could vary over time or it can be staritified for a better realization of the reality. The survey size could also vary such as

could defer this discussion to later, when actually trying some extensions

$n_A = 100$ or $n_A = 10000$ in later examples.

2.1.1 Markov chain Monte Carlo

The achievement of the simulation was done by Markov chain Monte Carlo (MCMC). The technique is a popular method to obtain posterior samples from distributions where analytic forms do not exist or are hard to be derived. **[Further explanation is going to be added.]** [attempt:brief explaination of MCMC and related topics]

2.1.2 Likelihood

There exist two data sets; survey data (n_A, x_A), and ambulance attended overdose data (U_t). The two data sets are simulated as follows. The true value of p_A was set as $p_A = 0.8$ for the survey data. It is assumed that the data was collected for a year (t=1,2,3, ..., 12) and x_t values were independentally generated from the Binomial distribution (2). ~~In~~ terms of overdose data, It is assumed that the true values of parameters for overdose model were $\mu = \log 0.05$, $\sigma = 1$. The vector of O_t was generated following the overdose model (3). The vector of U_t was generated from the Binomial relation of the two variables (4). The two generated vectors have the same length with the survey data (t=1,2,3, ..., 12). [Delete: Then the 12 sueveys were combined into a single survey since p_A is assumed fixed across time in this contexts.] Note that only U_t and x_t are known as the likelihood and p_A needs to be estimated first so as to estimate O_t which is the ultimate interest of the research.

n_A = ?

← the set-up described above has a single ×

2.1.3 Prior Distributions

Noninformative prior distributions are presumed as a start for simplicity.

$$p(p_A) \sim Beta(1,1) \text{ noninformative prior of ambulance model} \tag{5}$$

for

$$\left.\begin{aligned} \mu &\sim U(-10,0) \\ \sigma &\sim U(0,5) \end{aligned}\right\} \text{ noninformative prior of overdose model} \tag{6}$$

This leads the posterior distribution of variables of interest to heavily depend on the likelihood. Later, the noninformative priors will be changed and the impact of the changes over posteriors will be investigated.

2.2 Early Result

The result from the simple case scenario is illustrated below.

2.2.1 Posterior Distribution

Figure 2 is the boxplot of posterior samples of O_t. It is shown that our posterior estimates of O_t are fairly accurate since (1) the boxplots contain actual values of O_t within their interquartile range (IQR) and (2) the ranges of IQR and 95% range seem narrow. Notice that the range of the boxplot from a higher O_t values (t=1, 7, 11) is wider than the other ranges of the boxplots from smaller estimates of O_t (all t values but 1, 7, 11)

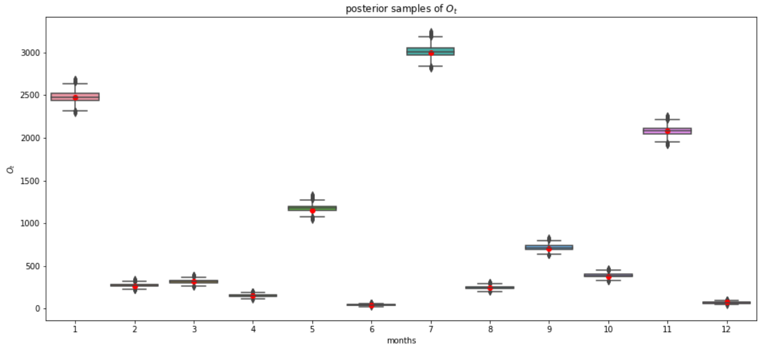


Figure 2: Boxplot of posterior samples of O_t (2000 samples for each month) with actual data points of simulated O_t values. The simulated values are shown as red dots.

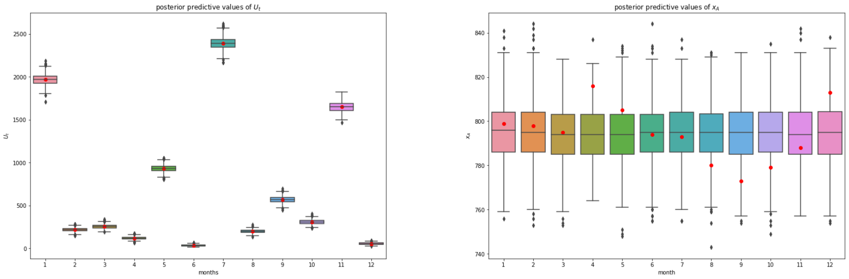
confusing to refer to O_t as data

also could give, in histogram form, posterior dist of $\sum_{i=1}^{12} O_i$

2.2.2 Posterior Predictive Check

Posterior predictive checking is a model validation technique that we simulate some replicated data under the fitted model then compare the new data to the observed data. If the model fits, then replicated data generated under the model should look similar to observed data. To put it another way, the observed data should look plausible under the posterior predictive distribution. This is really a self-consistency check: an observed discrepancy can be due to model misfit or chance. [5] [attempted: brief explannation of what is PPC]

Figure 3a is the boxplot of posterior predictive samples of U_t. It is shown that the posterior predictive estimates of U_t is failry accurate with the same two reasons regarding the accuracy of the posterior distribution of O_t. It is more obvious here that the range of the boxplots from higher O_t values (t=1, 7, 11) is wider than the other ranges of the boxplots from smaller estimates of O_t (all t values but 1,7, 11). Notice that the relative ranges of figure 3a follow the ones of figure 2; for months where U_t values are higher, the values of O_t are also higher than the average (t=1, 7, 11)



(a) ppc of U_t

(b) ppc of p_A

Figure 3: Boxplots of posterior predictive samples of x_A (1000 samples) with the actual data point of simulated U_t,x_A value. The simulated values are shown as a red dots.

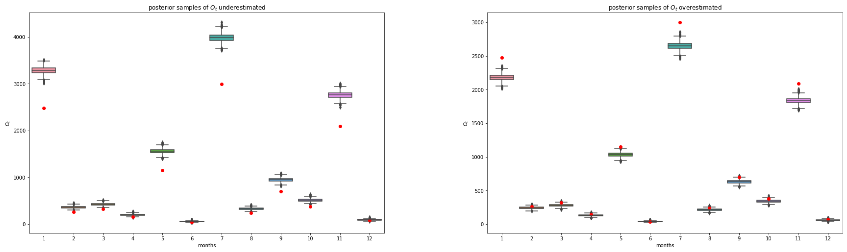
Figure 3b is the boxplots of posterior predictive samples of x_A. It is shown that the posterior predictive estimates of x_A is tolerably accurate since (1) the boxplots contain actual values of O_t within their lines connecting the maximum and the minimum and (2) the ranges of IQR seem modereately narrow.

2.3 Early Result: Contamination of p_A

One of the attention of this research project is to investigate how robust the model is from contaminations of the data sets. The first inspection is to check an impact of a contamination of p_A; what would happen if the estimation of p_A is biased? It is assumed that the survey data gives us a wrong estimate of p_A such that it would be underestimated or overestimated. We then want to see how the biased estimation of p_A affects the estimate of O_t, the total overdose.

Both of underestimation and overestimation were conducted for the analysis. In terms of underestimation, the simulated survey data (n_A, x_A) was generated with $p_A = 0.6$ while the true value of p_A is 0.8, and all the other assumptions hold the same. That is, x_A is generated from $x_A \sim Bin(n_A, 0.6)$, while U_t is generated from $U_t \sim Bin(O_t, 0.8)$ for every t. For overestimation, the simulated survey data was generated with $p_A = 0.9$ while the true value of p_A is 0.8, and all the other assumptions hold the same.

2.3.1 Posterior Distribution



(a) ppc of O_t: underestimated p_A

(b) ppc of O_t: overestimated p_A

Figure 4: Boxplot of posterior samples of O_t (1000 samples) where survey data is contaminated. The actual data point of simulated O_t values are shown as red dots.

Figure 4 is the boxplots of posterior samples where p_A is contaminated from the survey data. It is seen that the underestimation of p_A (4a) leads to an overestimation of O_t as the boxplots are above the red dots. This is justifiable considering the given data sets (likelihoods) and the relationship between the two models (4); given the likelihood of the ambulance attended overdoses, U_t, a value of O_t is an unknown parameter of the binomial distribution, $Bin(O_t, p_A)$. Let o_t, u_t are samples from O_t and the likelihood value of U_t respectively. Then o_t is proportional to the quantity generated by multiplying u_t and the inverse of $\hat{p}_A$ where $\hat{p}_A$ is underestimated estimator of p_A; hence $\frac{1}{\hat{p}_A}$ is overestimated which leads overestimation of O_t.

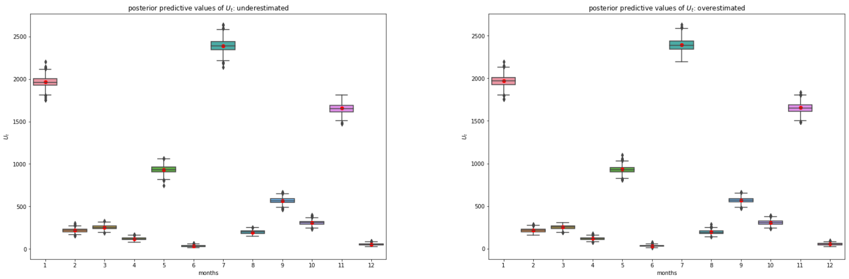
$$o_t \propto u_t \frac{1}{\hat{p}_A} \tag{7}$$

Figure 4b shows the opposite case. Overestimation of p_A leads underestimation of the inverse of p_A which causes underestimation of O_t. From both figures it is seen that the bias increases as the estimated values and the actual values get large.

2.3.2 Posterior Predictive Check

Figure 5 are the boxplots of posterior predictive samples where p_A is underestimated and overestimated respectively from the survey data. It is seen that none of the contaminations of p_A leads an contamination effect on U_t but only affects the O_t estimation.

The possible explanation is that the contaminations of p_A and O_t (from p_A) cancel out the bias component as combined so that U_t has no bias. Notice that U_t is a likelihood (data set) of O_t estimation. Hence it plays the role of y in Bayes theorem (1) and the likelihood, U_t, is used to obtain the biased posterior samples of O_t with the contaminated estimates of p_A. Here the pair of (O_t, P_A) are both biased and the pair is fitted by the likelihood, U_t. To obtain the posterior predictive samples, MCMC algorithm help obtain posterior samples of biased O_t and then the samples will be combined with biased p_A such that the new samples of $\tilde{U_t}$ from $p(\tilde{U_t}|U_t)$ is obtained. The result also matches with our intution. Posterior predictive checkes use the existing data points twice; firstly it uses the data to obtain the posterior samples of paramaters, and secondly the samples are used to produce posterior predictive samples. Given the fact that there could be an overfitting issue due to using the data twice, the result shown here seems reasonable.

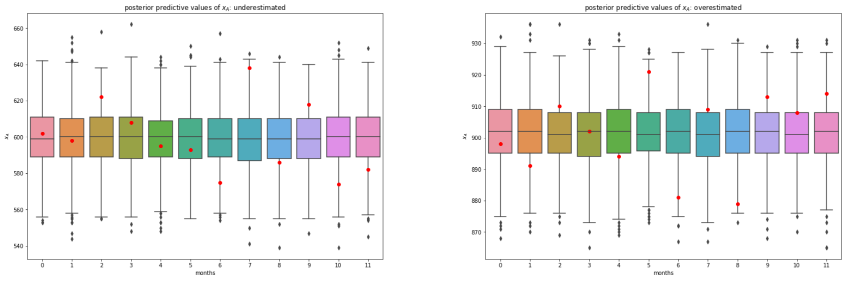


(a) ppc of U_t: underestimated p_A

(b) ppc of U_t: overestimated p_A

Figure 5: Boxplot of posterior samples of U_t (1000 samples) where survey data is contaminated. The actual data point of simulated U_t values are shown as red dots.

Figure 6 are the boxplots of posterior predictive samples of x_A where p_A is underestimated and overestimated respectively from the survey data. It is seen that none of the contaminations of p_A leads an effect of contamination on x_A; the actual simulated points (red dots) are close to the medians from the two boxplots. However, notice that the range of the estimated values are different between the two boxplots; the median from Figure 6a is around 600 whereas the median from Figure 6b is around 900. [???? I think this should go to the previous of the previous paragraph] This is because overdose model (3) does not affect the resulf from the ambulance model (2); only the ambulance model has an effect on the overdose model.



(a) ppc of x_A: underestimated p_A

(b) ppc of x_A: overestimated p_A

Figure 6: Boxplot of posterior samples of x_A (1000 samples) where survey data is contaminated. The actual data point of simulated x_A values are shown as red dots.

References

[1] BC Center for Disease Control.

[2] Cooper Smith. What is the opioid epidemic? url, 2019.

[3] Bayesian statistics. url, 2020.

[4] Buxton J Balshaw R Otterstatter M Macdougall L et al. Irvine MA, Kuo M. Modelling the combined impact of interventions in averting deaths during a synthetic-opioid overdose epidemic. *Addiction*, 2019.

[5] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis*. CRC Press, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, 2014.