ECO 517a Fall 2020

Lecture 5: Bayesian inference; mixed continuous and discrete random variables

Christopher A. Sims
Princeton University
sims@princeton.edu

September 14, 2020

The standard inference setup

- We assume we have a **model**, in the form of a conditional distribution of Y given a parameter vector β. The model implies a pdf $p(y \mid \beta)$ for $Y =\mid \beta$.

- We treat the model as known for sure. Uncertainty about the distribution of Y is entirely captured in the unknown β.

- Once we have seen the data Y, β is still unknown. $p(y \mid \beta)$ as a function of β, with y held fixed at its observed value, is called the **likelihood function**.

Bayesian additions: prior and posterior

- The Bayesian approach to inference aims at producing a probability distribution over the unknown β, conditional on the observed value of Y.

- As we have discussed, this can't be done without a marginal distribution for β. This distribution reflects what is known about β before seeing Y, and we will call its pdf $\pi(y)$.

- Then Bayes' rule tells us how to construct the conditional density $q(\beta \mid y)$ of β given y:
$$q(\beta \mid y) = \frac{\pi(\beta)p(y \mid \beta)}{\int \pi(\beta)p(y \mid \beta)\, d\beta}\,.$$

Example

- Suppose our model is that our two observations Y_1 and Y_2 are i.i.d., with a $\mathrm{Gamma}(2, a)$ distributiion. We don't know a.

- Suppose our prior on a is exponential, i.e. $\pi(a) = e^{-a}$.

- Then our posterior pdf for a is proportional to

$$e^{-a}a^4 y_1 y_2 e^{-a(y_1+y_2)} = a^4 y_1 y_2 e^{-a(1+y_1+y_2)} \; .$$

- This is not the posterior pdf, because we haven't calculated the denominator of Bayes' rule, which would scale it to integrate to one.

Example.2

- However, here, as often happens, we can avoid the calculus by noting that, as a function of a, this expression is proportional to a $\mathrm{Gamma}(5, 1+y_1+y_2)$ pdf. When properly normalized, therefore, it would just be this Gamma pdf.

- In this example, the likelihood itself is integrable, and indeed proportional to a $\mathrm{Gamma}(5, y_1+y_2)$ pdf.

- The likelihood, normalized to integrate to one, is the limiting form of the posterior distribution if we had a prior pdf of the form be^{-ba} and let b go to zero.

- Often in practice it is convenient to treat the likelihood as proportional to the posterior density. This is called using a “flat prior”.

Unbiased estimation

- Suppose we are interested in $2/a$, which is the expected value of Y_i in this example.

- The expectation of $2/a$ given the data is $(y_1 + y_2 + 1)/2$ or, with a flat prior, $(y_1 + y_2)/2$. (You should be sure you can verify this.)

- For the flat prior posterior mean of $2/a$, which turns out to be just the sample mean of the y_i's, we can say that **before** we saw the data, the expected value of $(Y_1 + Y_2)/2$ was $2/a$, which is the definition of an **unbiased** estimator of $2/a$. We'll discuss unbiasedness, and estimation in general, in more detail later.

Mixing discrete and continuous randomness

- When a random variable Y with probability 1 takes on only a finite number of values, the probability of each of its possible values is a function of y that is often called a “probability mass function” (pmf), to distinguish it from a “probability density function” (pdf) that characterizes the distribution of a continuously distributed variable.

- We will still call this a pdf, not a pmf, and this has a firm mathematical foundation.

Discrete observations, with a continuously distributed parameter

- One common situation where we have to think about discrete and continuous random variables jointly arises when a discrete variable Y has a distribution with a continuous parameter β.

- For example, suppose we are interested in inference about the population value of the probability p_0 that a randomly drawn individual will have 0 years of education.

Likelihood

- In the AK data set there will be a number n with 0 years of education and $N - n$ with more education. We can do our inference for p_0 by considering the random variable Y that is 0 when the sampled individual has some schooling, and 1 when he has none.

- The observed sample then is a sequence of N observations of this random variable, with some 1's and many 0's. The probabiity of observing the sample we have is

$$\prod_{i=1}^{N} p_0^{Y_i}(1-p_0)^{1-Y_i} = p_0^n(1-p_0)^{N-n} \, .$$

If we are interested in inference about p_0, this is the likelihood function.

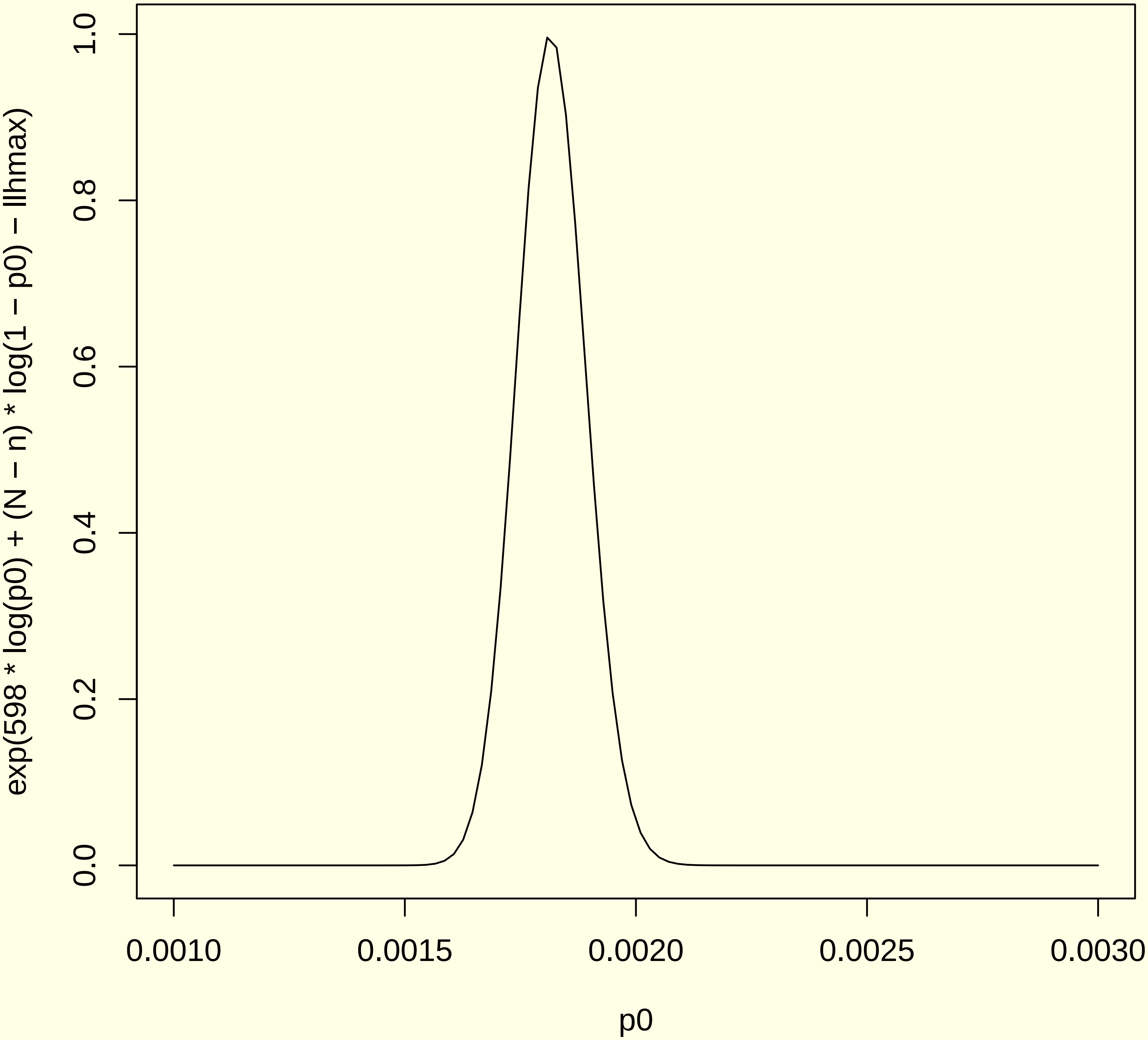
Prior, posterior

- Here it might seem "unprejudiced" to just let the prior distribution on p_0 be flat on (0,1).

- Then by Bayes' rule the posterior density on p_0 is proportional to the likelihood, which in turn we can recognize as proportional to a $\mathrm{Beta}(n+1, N-n+1)$ pdf.

- A plot of the posterior density is on the next page. It shows that p_0 is with high probability between about .16% and .20%. The proportion of the sample with no education is .1815%.

- The mode of the posterior pdf is just the sample proportion, while its posterior expectation is slightly different: .1817%.


exp(598 * log(p0) + (N − n) * log(1 − p0) − llhmax)
0.0
0.2
0.4
0.6
0.8
1.0
0.0010
0.0015
0.0020
0.0025
0.0030
p0

Continuously distributed observation, discrete parameter

- To keep this simple, we will make it unrealistic.

- Suppose we are considering two possible models for the distribution of `logwage` in the AK dataset: an exponential pdf ae^{-ax} or a Pareto: $a/(1+ax)^2$.

- Both of these assume x is non-negative, while the actual `logwage` data includes 195 negative observations. We'll just omit those, pretending the remaining 99.94% of the sample is the whole sample.

- A negative value for the log of weekly wages implies an annual income of less than $\$52$.

`logwage` example continued

- We could treat the a parameter in each model as unknown, which would be more realistic, but we'll discuss later how to do that in this situation of two models.

- Instead we will assume we know that in the exponential model $a = .1695$ and in the Pareto model $a = 3.0634$. These values have been estimated from the data, but we'll proceed as if we knew them exactly.

- There is only one thing unknown here: which model is correct. We can treat that as a discrete parameter taking values $\mu = 1$ (exponential) and $\mu = 0$ (Pareto).

- The likelihood function with the data values held fixed and the unknown parameter μ varying, takes on just two values:

$$\begin{aligned} \mu = 0: & \qquad \frac{a_p^N}{\prod_i (1 + a_p x_i)^2} \\ \mu = 1: & \qquad a_e^N \exp\Bigl(-a_e \sum_i x_i\Bigr) \end{aligned}$$

- We can evaluate these expressions. $\mu = 0$ gives us -1,568,866 for the log likelihood, while $\mu = 1$ gives us -914,402.9.

- The difference in these log likelihoods, exponentiated, gives the odds ratios on the two models. The ratio favoring $\mu = 1$ is infinite to machine precision.

Discussion of the two examples

- In both these cases we have treated the probabilities on discrete points the same way we treat pdf values in forming likelihoods and applying Bayes' rule. The only difference is that in normalizing prior times likelihood to integrate to one, we sum when the prior times likelihood is discrete and integrate when it is continuous.

- In the second case, we arrived at a common type of result: When comparing discrete models, odds ratios often turn out to be extreme.

- It's seldom true that the array of models we consider covers all possible models, or that there are no models "in between" the models we consider.

- Extreme odds ratios across models are usually a signal that we should consider a richer array of models, particularly other models "close" to the one with highest posterior odds.

Mixed discrete and continuous distributions

- Occasionally in practice we need to deal with cases where every real number (or element of $\mathbb{R}^k$) is possible, but instead of their all having zero probability as individual points, some have positive probability.

- For example, in the AK data, `logwage` has many observations that, when exponentiated, multiplied by 52, and decreased by 5.0, are exact multiples of 5000. That is, it looks like many observations generated a "weekly wage" by dividing a round-number annual income by 52. The number of people in the sample reporting a `logwage` corresponding to a 20000 annual income is 11617.

lumpiness in the `logwage` distribution

- It appears that the log weekly wage data was generated by asking for annual incomes, adding 5, dividing by 52.

- Why adding 5? Because the result has to be logged, and there were 25 people with incomes of zero. Adding 5 eliminated the "log of zero" error messages.

- There are values other than the round numbers. Within the interval corresponding to annual incomes of 19895 and 20095, there are 1314 that are not equal to that corresponding to 20000 (besides the 11617 that do).

A model for the `logwage` distribution

- A serious model for this rounding error would take us too far afield, but a simple one might look like this.

- Put probability q_i on observations corresponding to annual incomes of $10000i$, $i = 1, \dots, 20)$, with the sum of the q_i's $\alpha < 1$. Then in addition there is a probability $1 - \alpha$ of the observation being drawn from, say an exponential pdf ae^{-ax}.

- Then it might seem reasonable, and it is, to form the likelihood by taking the "pdf" value for observations that hit one of the round numbers to be the corresponding q_i, and taking the pdf value corresponding to other observations to be $(1 - \alpha)ae^{-ax}$.

- The likelihood for the full sample than becomes

$$\prod_{i=1}^{20} q_i^{n_i} \cdot ((1-\alpha)a)^{N-\sum n_i} \exp\Bigl(-a \sum_{x_j \in A} x_j\Bigr) \, .$$

  where n_i is the number of observations with `logwage` matching the $10000i$ annual income and A is the set of `logwage` values that do not match any $10000i$ values.

- We could use this likelihood, together with a prior, to construct a posterior joint distribution for a and $\{q_i\}_{i=1}^{20}$.

How to justify this sensible procedure

- When we integrate an ordinary pdf we are integrating with respect to “Lebesgue measure”, which assigns sizes to sets in $\mathbb{R}^k$ in the conventional way (length, area, volume).

- When we sum probabilities of points in a discrete distribution, we are integrating again, but here with respect to “counting measure”, which assigns sizes to sets by counting the points in the set.

- In a mixed discrete-continuous distribution, we can take the measure with respect to which we are integrating to be the sum of Lebesgue measure and counting measure attached to the points with discrete weight. Then to find the probability of a set, we sum the probabilities of the points in it with non-zero probability, and add to that the integral of the pdf for the continuous part. This is integration with respect to a pdf on the sum of the counting and Lebesgue measures.

- Bayes' rule applies to densities with respect to these mixed measures just as it does to continuous densities. We just have to remember what the integral sign means in these cases.

Limits to the simple form of Bayes' rule

- In the example we considered of a mixed continuous-discrete random variable, the points in x-space that had non-zero probability did not depend on an unknown parameter. If they did, the simple form of Bayes' rule based on pdf's no longer would work.

- Here's an example where it would not. Suppose you had a sample of wage observations that you believe are drawn from a distribution with a minimum wage $\bar{w}$ which people earn with probability q, and that people not earning the minimum wage have an exponential distribution of wages with pdf $ae^{a(\bar{w}-w)}$ over $(\bar{w}, \infty)$.

Details on the example

- The unknown parameters here are q, $\bar{w}$ and a, but even with convenient forms for the prior on these parameters, formulating Bayes'rule with densities does not work here, even though Bayesian inference is possible.

- One way to see the problem is to note that if we take the model seriously, as soon as we have seen two values of w in the sample that exactly match, we know that must be $\bar{w}$, with no uncertainty.

- If you want a challenge, work out the posterior distribution on $\bar{w}$ and a in a sample that has no repeated values for w, assuming you know q exactly a priori and you have a prior pdf $\bar{w}e^{-\bar{w}}$ on $\bar{w}$ and e^{-a} on a, with the priors independent.