Example sheet 2

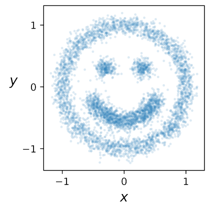
Bayesian inference

Data Science—DJW—2023/2024

Questions labelled * are more challenging. Some of the questions ask for pseudocode; you are encouraged to actually implement your answers! You can test your code using the online tester. There is a notebook with templates for answers and instructions for submission on the course materials webpage.

Following this example sheet is a page with hints for each question. There are also a set of supplementary questions. These are not intended for supervision (unless your supervisor directs you otherwise).

Question 1. Define a function `rxy()` that produces a random pair of values (X, Y) which, when shown in a scatterplot, produces a smiley face like this. Also plot the marginal distributions of X and Y.



Question 2. Consider this code for generating random variables X and Y:

```
x = np.random.uniform()
y = np.random.geometric(p=x)
```

Derive the marginal likelihood $\Pr_Y(y)$, and the conditional likelihood $\Pr_X(x \mid Y = y)$.

Question 3. I have a dataset $x_1, \dots, x_n$ which I model as samples from $\text{Poisson}(\theta)$. The parameter θ is unknown, and I shall use $\Theta \sim \text{Exp}(1)$ as my prior. Give pseudocode to plot the posterior distribution of Θ and to compute a 95% posterior confidence interval. *[Optional: To test your code using the online tester, fill in the answer template for `exp_poisson_confint`.]*

Question 4. I have a dataset $x_1, \dots, x_n$, which I model[1] as samples from $\text{Uniform}[a, a+b]$. The parameters a and b are unknown, and I shall use $A \sim \text{Exp}(\lambda_0)$ and $B \sim \text{Exp}(\mu_0)$ as my prior. Give pseudocode to plot the posterior distribution for B and to compute a 95% posterior confidence interval for it. *[Optional: To test your code using the online tester, fill in the answer template for `exp_uniform_confint`.]*

Question 5. I have a dataset of monthly average temperatures in Cambridge from 2010 onwards, and I propose the model

$$\text{Temp} \sim \alpha + 6.6826 \sin\big(2\pi(\mathtt{t} - 0.27731)\big) + \gamma(t - 2000) + N(0, 1.4183^2)$$

where α and γ are unknown. Using prior distribution $\alpha \sim N(10, 5^2)$ and $\gamma \sim N(0, 0.1^2)$ …

(a) Give pseudocode to compute a 95% confidence interval for γ.

(b) The predicted average temperature in 2050 is $\text{pred}(2050) = \alpha + 50\gamma$. Give pseudocode to compute a 95% confidence interval for pred(2050).

[Optional: To test your code using the online tester, fill in the answer templates for `climate_inc_confint` and `climate_pred_confint`.]

[1] An earlier version of this example sheet mistakenly used $\text{Uniform}[a, b]$. Corrected 2023-10-23.

Question 6. I sample $x_1, \dots, x_n$ from Uniform$[0, \theta]$. The parameter θ is unknown, and I shall use $\Theta \sim$ Pareto(b_0, α_0) as my prior, where $b_0 > 0$ and $\alpha_0 > 1$ are known. This has the cumulative distribution function

$$\mathbb{P}(\Theta \leq \theta) = \begin{cases} 1 - \big(b_0/\theta\big)^{\alpha_0} & \text{if } \theta \geq b_0, \\ 0 & \text{if } \theta < b_0. \end{cases}$$

(a) Calculate the prior likelihood for Θ.

(b) Show that the posterior distribution of $(\Theta \mid x_1, \dots, x_n)$ is Pareto, and find its parameters.

(c) Find a 95% posterior confidence interval for Θ.

(d) Find a different 95% posterior confidence interval. Which is better? Why?

Question 7. I have a collection of numbers $x_1, \dots, x_n$ which I take to be independent samples from the Normal(μ, σ_0^2) distribution. Here σ_0 is known, and μ is unknown. Using the prior distribution $M \sim$ Normal(μ_0, ρ_0^2) for μ, show that the posterior density is

$$\Pr_M(\mu \mid x_1, \dots, x_n) = \kappa e^{-(\mu - c)^2/2\tau^2}$$

where κ is a normalizing constant, and where you should find formulae for c and τ in terms of σ_0, μ_0, and ρ_0, and the x_i. Hence deduce that the posterior distribution is Normal(c, τ^2). *[Note: 'M' is the upper-case form of the Greek letter 'μ'.]*

Question 8. I have a collection of numbers

$$[4.3, 2.8, 3.9, 4.1, 9, 4.5, 3.3]$$

which look like they mostly come from a Gaussian distribution, but with the occasional outlier. I'll model the data as

$$X \text{ is } \begin{cases} \text{Normal}(\mu, 0.5^2) & \text{with probability 99\%} \\ \text{Cauchy} & \text{with probability 1\%} \end{cases}$$

where μ is unknown. The likelihood function for a single datapoint is

$$\Pr_X(x|\mu) = 0.99\, \text{pdf}_{N(\mu,0.5^2)}(x) + 0.01\, \text{pdf}_{\text{Cauchy}}(x)$$

where pdf_N and $\text{pdf}_{\text{Cauchy}}$ are the pdfs for Normal and Cauchy random variables respectively. *[Note. The Cauchy random variable occasionally generates wildly huge values, which makes it a good model for big outliers. The library function* `scipy.stats.cauchy.pdf(x)` *computes its pdf.]*

Using a Normal$(0, 5^2)$ prior distribution for μ, give pseudocode to plot its posterior distribution.

Question 9*. Consider the outlier model from question 8. How likely is it that the datapoint with value 9 is an outlier?

Question 10*. I am prototyping a diagnostic test for a disease. In healthy patients, the test result is Normal$(0, 2.1^2)$. In sick patients it is Normal$(\mu, 3.2^2)$, but I have not yet established a firm value for μ. In order to estimate μ, I trialled the test on 30 patients whom I know to be sick, and the mean test result was 10.3. I subsequently apply the test to a new patient, and get the answer 8.8. I wish to know whether this new patient is healthy or sick.

(a) In this question there are two unknown quantities: μ, and $h \in \{\text{healthy}, \text{sick}\}$ the status of the new patient. Model the former as a random variable M with prior distribution Normal$(5, 3^2)$ and the latter as a random variable H with prior distribution

$$\Pr_H(h) = 0.99 \times 1_{h=\text{healthy}} + 0.01 \times 1_{h=\text{sick}}.$$

Write down the joint prior likelihood for (M, H).

(b) In this question the data consists of 31 values, test results $x_1, \dots, x_{30}$ from the known sick patients and test result y from the new patient. Write down the data likelihood $\Pr(x_1, \dots, x_{30}, y|\mu, h)$.

(c) Find the posterior density of (M, H). Leave your answer as an unnormalized density function. It should simplify to be a function of $\bar{x}$ and y, where $\bar{x}$ is the mean test result for the known sick patients.

(d) Give pseudocode to compute the posterior distribution of H, i.e. compute $\mathbb{P}(H = h \mid \text{data})$ for both $h = \text{healthy}$ and $h = \text{sick}$.

Hints and comments

Question 1. Try extending the Gaussian mixture model from section 1. For plotting, here's some code. It assumes that you have stored your samples in a numpy array of shape $n \times 2$, one row per sample point, columns for x and y.

```
fig,((ax_x,dummy),(ax_xy,ax_y)) = plt.subplots(2,2, figsize=(4,4),
    sharex='col', sharey='row', gridspec_kw='height_ratios':[1,2], 'width_ratios':[2,1])
dummy.remove()
ax_xy.scatter(xy[:,0], xy[:,1], s=3, alpha=.1)
ax_x.hist(???, density=True, bins=60) # fill in the ???
ax_y.hist(???, density=True, bins=60, orientation='horizontal') # fill in the ???
plt.show()
```

Question 2. There are two versions of the Geometric distribution; look up the numpy help page to see which one is being used here. For the marginal likelihood, write out the joint likelihood and integrate. For the conditional likelihood, the calculation is similar to exercise 5.2.2 from lecture notes.

Question 3. All Bayesianist computations start in exactly the same way. Before you start, write out the likelihood of the data, $\Pr(x_1, \ldots, x_n \mid \Theta = \theta)$. Then, (1) take a sample $\theta_1, \ldots, \theta_m$ from the prior distribution, (2) compute weights by evaluating the likelihood of the data at each of these samples θ values, and rescale so that weights sum to one.

For plotting the posterior distribution, use `plt.hist` as described in section 6.2 of notes. For computing a 95% confidence interval, see section 8.4.

Question 4. Use the same general approach as question 3. But now there are two unknown parameters, so take a sample $(a_1, b_1), \ldots, (a_m, b_m)$ from the joint prior distribution (even though it's only the b samples that we want to report at the end). See exercise 8.3.2 from lecture notes for more hints.

Question 5. Part (a) is exactly like question 4: there are two unknown parameters, and we want to report the posterior distribution of one of them. There is more data here, though, which makes the data likelihood very small, and to avoid underflow you should use the log-sum-exp trick described in exercise 8.3.4 in lecture notes. Also, you'll have to think about how to vectorize your code efficiently for it to run fast enough.

For part (b): Each sample (α_j, γ_j) of the unknown parameters yields a sample $\text{pred}_j(2050) = \alpha_j + 50\gamma_j$. Use these samples, together with the weights you computed earlier, to compute a 95% confidence interval for pred(2050). The thinking behind this is discussed in section 8.5 of lecture notes.

Question 6. For part (a), remember from IA Probability that the pdf (i.e. the likelihood) is just the derivative of the cdf. The question tells us the cdf—just differentiate it! Write it out carefully using indicator function notation, $1_{\theta \geq b_0}$. This is often a good idea, when we're working with parameters that affect boundaries (as in lecture notes example 1.3.6).

For the rest: **all Bayesian calculations start in exactly the same way.** Before you start, write out the likelihood of the observed data $\Pr(x_1, \ldots, x_n \mid \Theta = \theta)$. Then (1) write down the prior likelihood $\Pr_\Theta(\theta)$, (2) apply Bayes's rule which says that the posterior likelihood is

$$\Pr_\Theta(\theta \mid x_1, \ldots, x_n) = \kappa \Pr_\Theta(\theta) \Pr(x_1, \ldots, x_n \mid \Theta = \theta).$$

In this question, write out the likelihood of the data using indicator notation, as in example sheet 1 question 3. Once you have the posterior density, gather together the θ terms, and you should end up with the density of another Pareto.

For the posterior confidence interval: the definition of a posterior confidence interval is in lecture notes section 8.4. You just have to solve the equations for `lo` and `hi`, using the cumulative distribution function for the Pareto.

Question 7. All Bayesian calculations start in exactly the same way. Before you start, write out the likelihood of the observed data $\Pr(x_1, \dots, x_n \mid M = \mu)$. Then (1) write down the prior likelihood $\Pr_M(\mu)$, (2) apply Bayes's rule which says that the posterior likelihood is

$$\Pr_M(\mu \mid x_1, \dots, x_n) = \kappa \Pr_M(\mu) \Pr(x_1, \dots, x_n \mid M = \mu).$$

Remember, this is a density function for a random variable M, and the argument is μ. Write your answer to gather together all the μ terms as much as you can. This involves expanding quadratic terms and completing the square. Any terms that don't involve μ can be amalgamated with the constant factor κ. What you end up with should look like a Normal density function, as a function of μ, and this lets you conclude that the posterior distribution is Normal.

When a question asks "find the posterior distribution", you should start by calculating the posterior density, leaving it unnormalized i.e. including a constant factor, call it κ. Then (a) if you recognize this as a standard density function, as is case here, just give its name; (b) if it's easy to find κ using "densities sum to one" then do so; (c) otherwise leave your answer as an unnormalized density function.

Question 8. This is nearly exactly the same as question 3. The only difference is the formula for the likelihood of the data, $\Pr(x_1, \dots, x_n \mid M = \mu)$.

It's a good exercise to derive the likelihood function $\Pr_X(x)$. The calculation is very similar to finding the likelihood for the Gaussian mixture model, example 5.3.5 from lecture notes.

Question 9. There are two unknowns in this question: the unknown parameter μ, and the unknown of whether the value $x^* = 9$ is an outlier. Write $k = 1_{x^* \text{ is outlier}}$ for this second unknown. Either $k = 0$ (it's not an outlier), or $k = 1$ (it is an outlier).

The likelihood function for the other datapoints is $\Pr_X(x|\mu)$ from question 8. The likelihood for the datapoint we're interested in is

$$\Pr_{X^*}(x^*) = \begin{cases} \text{pdf}_{N(\mu, 0.5^2)}(x^*) & \text{if } k = 0 \\ \text{pdf}_{\text{Cauchy}} & \text{if } k = 1. \end{cases}$$

The likelihood of the dataset is the likelihood of all the datapoints, both the other datapoints and also x^*.

Once we've got a formula for the likelihood of the datapoints, we simply use the standard computational Bayes procedure, in this case for a problem with two unknown parameters (like question 4). This question asks for the probability that x^* is an outlier, i.e. the posterior probability that $k = 1$. Section 6.2 tells us how to approximate posterior probabilities. You should get an answer very close to 1.0.

Question 10. This is another question with two unknowns, like questions 9 and 4. For part (b), for the likelihood $\Pr_Y(y \mid \mu, h)$, see the Gaussian mixture model in exercise 5.3.5 in lecture notes. For part (c), your formula for the posterior distribution will involve equations very similar to question 7.

Supplementary question sheet 2

Bayesian inference

These questions are not intended for supervision (unless your supervisor directs you otherwise). Some of require careful maths, some are best answered with coding, some are philosophical.

Question 11. Consider this code for generating random variables $X \to Y \to Z$:

```
x = np.random.uniform()
y = np.random.binomial(n=1, p=x)
z = np.random.normal(loc=y, scale=ε)
```

Show that

$$\Pr_Y(1 \mid X = x, Z = z) = \frac{x}{x + (1-x)e^{(1-2z)/2\varepsilon^2}}.$$

How does $\Pr_Y(1 \mid X = x, Z = z)$ depend on x and z when $\varepsilon \approx 0$? What if ε is very large?

[If we want to predict Y, and we have x and z available, should we use $\Pr_Y(y \mid X = x, Z = z)$, or $\Pr_Y(y \mid X = x)$, or $\Pr_Y(y \mid Z = z)$? The obvious answer is that we should use the first, since it uses all available data.

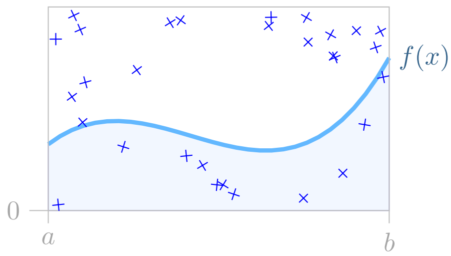
[But suppose we're interested in predicting Y, and we've trained a predictor on (x, y, z) data generated according to the code above, but in deployment the data comes from a slightly different model – which of the three predictors is robust to this change in environment? If the first line of code is different for the new data environment, then the first and second predictors still work correctly. If the second line of code is different, then all bets are off. If the third line of code is different, only the second predictor still works. So, for robust prediction, we might prefer the second predictor. It's called the 'causal predictor' since it only uses the input variable that directly causes the response we're interested in.

[The challenge is that, in typical machine learning tasks, we don't know which of our predictor variables are causal and which aren't.]

Question 12. Suppose we're given a function $f(x) \geq 0$ and we want to evaluate

$$\int_{x=a}^{b} f(x)\, dx.$$

Here's an approximation method: (i) draw a box that contains $f(x)$ over the range $x \in [a, b]$, (ii) scatter points uniformly at random in this box, (iii) return $A \times p$ where A is the area of the box and p is the fraction of points that are under the curve. Explain why this is a special case of Monte Carlo integration.



Do NOT give a wishy-washy qualitative argument along the lines of "there are random points, and we're evaluating an integral, so it's a type of Monte Carlo". Monte Carlo has a precise meaning: $\mathbb{E}\, h(X) \approx n^{-1} \sum_i h(x_i)$. In your answer you should (a) explain the random variable in question, (b) specify the h function, (c) give an explanation along the lines of section 5.1 of lecture notes.

Question 13 (Leaky priors). I repeatedly attempt a task, and each time I attempt it I succeed with probability θ and fail with probability $1 - \theta$. The parameter θ is unknown, so I model it as a random variable Θ. Ever the optimist, my prior for Θ is heavily biased in favour of large values for θ:

$$\Pr_\Theta(\theta) = \varepsilon 1_{\theta \leq 1/2} + (2 - \varepsilon) 1_{\theta > 1/2}$$

for some known small value $\varepsilon > 0$; this implies $\mathbb{P}(\Theta \leq 1/2) = \varepsilon/2$.

But I experience an unbroken run of n failures. How big does n need to be, for me to concede there's a 50% posterior probability that $\Theta \leq 1/2$? How big would it need to be, if $\varepsilon = 0$?

Question 14. In lectures we investigated a dataset of police stop-and-search actions. Let the outcome for record i be $y_i \in \{0,1\}$, where 1 denotes that the police found something and 0 denotes that they found nothing. Consider the probability model $Y_i \sim \text{Binom}(1, \beta_{\texttt{eth}_i})$ where $\texttt{eth}_i$ is the recorded ethnicity for the individual involved in record i, and where the parameters β_{As}, β_{Blk}, β_{Mix}, β_{Oth}, β_{Wh} are unknown. As a prior distribution, suppose that the five β parameters are all independent $\text{Beta}(1/2, 1/2)$ random variables.

(a) Write down the joint prior density for $(\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$.

(b) Find the joint posterior distribution of $(\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$ given the y data.

Question 15 (Sequential Bayes). I have a biased coin, with unknown probability of heads θ. I toss it n times, with outcomes $x_1, x_2, \dots, x_n$ where $x_n = 1$ indicates heads and $x_n = 0$ indicates tails. My prior belief is $\Theta \sim \text{Uniform}[0,1]$. Here are two approaches to applying Bayes's rule:

- *One-shot Bayes.* Use Bayes's rule to compute the posterior of Θ, given data $(x_1, \dots, x_n)$, using prior $\Theta \sim \text{Uniform}[0,1]$, and assumimg that coin tosses are independent.
- *Sequential Bayes.* Use Bayes's rule to compute the posterior of Θ given data x_1, using the uniform prior; let the posterior density be $p_1(\theta)$. Apply Bayes's rule again to compute the posterior of Θ given data x_2, but this time using $p_1(\theta)$ as the prior; let the posterior density be $p_2(\theta)$. Continue applying Bayes's rule in this way, until we have found $p_n(\theta)$.

State the posterior distribution found by one-shot Bayes. Prove by induction on n that sequential Bayes gives the same answer.

Sequential Bayes and one-shot Bayes give the same answer for any inference problem, not just this coin-tossing example. Can you prove the general case?

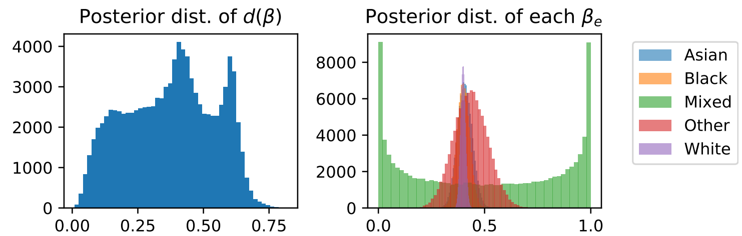
Question 16. In the setting of question 14, I wish to measure the amount of police bias. Given a 5-tuple of parameters $\beta = (\beta_{\text{As}}, \beta_{\text{Blk}}, \beta_{\text{Mix}}, \beta_{\text{Oth}}, \beta_{\text{Wh}})$, I define the overall bias score to be

$$d(\beta) = \max_{e,e'} \left| \beta_e - \beta_{e'} \right|.$$

If $d(\beta)$ is large, then there is *some* pair of ethnicities with very unequal treatment.

As a Bayesian I view β as a random variable taking values in $[0,1]^5$, therefore $d(\beta)$ is a random variable also. To investigate its distribution, I sample β from the posterior distribution that I found in question 14, I compute $d(\beta)$, and I plot a histogram. The output, shown on the left, is bizarre. To help me understand what's going on, I plot histograms of each of the individual β_e coefficients, shown on the right.

Explain the results. *[Hint. Explore the Beta distribution numerically. For what parameters does it have a bimodal distribution? What are the posterior distributions in this question?]*



Question 17. I have a coin, which might be biased. I toss it n times and get x heads.

I am uncertain whether or not the coin is biased. Let $m \in \{\text{fair}, \text{biased}\}$ indicate which of the two cases is correct; and if it is biased let θ be the probability of heads. The probabilty of observing x heads is thus

$$\Pr(x \mid m, \theta) = \begin{cases} \binom{n}{x} \theta^x (1-\theta)^{n-x} & \text{if } m = \text{biased} \\ \binom{n}{x} (1/2)^x (1-1/2)^{n-x} & \text{if } m = \text{unbiased} \end{cases}$$

As a Bayesian I shall represent my uncertainty about m with a prior distribution, $\Pr_M(\text{fair}) = p$, $\Pr_M(\text{biased}) = 1 - p$. If it is biased, my prior belief is that the probability of heads is $\Theta \sim \text{Uniform}[0,1]$.

(a) Write down the prior distribution for the pair (M, Θ), assuming independence as usual.

(b) Find the posterior distribution of (M, Θ) given x.

(c) Find $\mathbb{P}(M = \text{unbiased} \mid x)$, i.e. the posterior probability that the coin is unbiased.

This is a Bayesian question, and it's answered in the same way as any other Bayesian question: write down the prior density $\Pr_{M,\Theta}(m,\theta)$, *write down the data density* $\Pr(x \mid m,\theta)$, *and multiply them together (times a constant factor) to get the posterior* $\Pr_{M,\Theta}(m,\theta \mid x)$. *To keep track of all the cases, it may be helpful to use indicator functions, both for* $\Pr_M$ *and for* $\Pr(x \mid m,\theta)$.

Part (c) is about nuisance parameters, as in exercise 7.4 in lecture notes (look at the mathematical solution of that exercise). Once we've found the posterior density, say $\Pr_{M,\Theta}(m,\theta) = \kappa f(m,\theta)$ *where* κ *is the normalizing constant, we have to integrate out* θ *to find the marginal distribution, as in exercise 7.4:*

$$\mathbb{P}(M = fair \mid x) = \int_\theta \kappa f(fair,\theta)\,d\theta \qquad \mathbb{P}(M = biased \mid x) = \int_\theta \kappa f(biased,\theta)\,d\theta.$$

Then solve for κ, *using the "densities sum to one" rule, as in exercise 7.5 from lecture notes.*

This question is an illustration of Bayesian model selection, which you can read about in section 7.4 of lecture notes.

Question 18. (a) Suppose we have a single observation x, drawn from Normal$(\mu+\nu,\sigma^2)$, where μ and ν are unknown parameters, and σ^2 is known. Explain why the maximum likelihood estimates for μ and ν are non-identifiable.

(b) For μ use Normal(μ_0,ρ_0^2) as prior, and for ν use Normal(ν_0,ρ_0^2), where μ_0, ν_0, and ρ_0 are known. Find the posterior density of (μ,ν). Calculate the parameter values $(\hat\mu,\hat\nu)$ where the posterior density is maximum. (These are called *maximum a posteriori estimates* or *MAP estimates*.)

(c) An engineer friend tells you "Bayesianism is the Apple of inference. You just work out the posterior, and everything Just Works™, and you don't need to worry about irritating things like non-identifiability." What do you think?

Question 19. Here's my answer to question 1:

```
k = np.random.choice(4, p=[.6,.3,.05,.05], size=n)
t = np.random.uniform(size=n)
x = np.column_stack([np.sinπ(2**t), 0.55*np.sinπ(2**(0.4*t+0.3)), -0.3*np.ones(n), 0.3*np.ones(n)])
y = np.column_stack([np.cosπ(2**t), 0.55*np.cosπ(2**(0.4*t+0.3)), 0.3*np.ones(n), 0.3*np.ones(n)])
xy = np.column_stack([x[np.arange(n), k], y[np.arange(n), k]])
xy = np.random.normal(loc=xy, scale=.08)
```

Compute the distribution of $(X \mid Y = 0.3)$. Give your answer as a histogram.

You will need to derive your own method for sampling, along the lines of the derivation of computational Bayes in section 5.2. The difference here is that instead of using Bayes's rule

$$\Pr_X(x \mid Y=y) = \kappa \Pr_{X,Y}(x,y) = \kappa \Pr_X(x) \Pr_Y(y \mid X=x)$$

you will need to use a version more suited to the generation method used here,

$$Pr_{X,Y}(x,y) = \sum_k \int_t \Pr(x,y,k,t)\,dt = \sum_k \int_t \Pr_K(k)\Pr_T(t)\Pr_X(x \mid k,t)\Pr_Y(y \mid k,t)\,dt\,.$$

You should end up with a Monte Carlo integration that uses (K,T,X) *samples.*