Prof. David Draper
Department of Statistics
Baskin School of Engineering
University of California, Santa Cruz
Winter 2022

STAT 206 (Applied Bayesian Statistics)

Take-Home Test 3: Part 1 (*Revised* 17 Mar 2022)

Absolute due date: Uploaded to `canvas.ucsc.edu` by 11.59pm on **20 Mar 2022**

Name: Kevin Guillen

Here are the ***revised*** ground rules: this test is open-book and open-notes, and has three parts. Part 1 consists of 7 true/false questions, each worth 10 points, for a total of 70 points; this part is ***mandatory for all STAT 206 students***. Part 2 has a single calculation question in it, worth 220 points; this part is also ***mandatory for all STAT 206 students***. Part 3 is entirely ***optional for all STAT 206 students*** and acts as a source of ***extra credit*** (up to 220 additional points): any points earned here will be added to the numerator, but not the denominator, in computing your course percentage correct

(total points achieved) / (total points assigned) .

Undergraduates who wish to gain full mastery of all of the material presented this quarter are strongly encouraged to participate in office hour sessions from now through Sun 20 Mar 2022.

Some advice on style as you write up your solutions: pretend that you're sitting next to the grader, having a conversation about problem (x) part (y). You say, "The answer is z," and the grader says, "Why?" You then give your explanation, as succinctly as possible to get your idea across. The right answer with no reasoning to support it, or incorrect reasoning, will get **half credit**, so try to make a serious effort on each part of each problem (this will ensure you at least half credit). In an AMS graduate class I taught in 2012, on a take-home test like this one there were 15 true/false questions, worth a total of 150 points; one student got a score of 92 out of 150 (61%, a D−, in a graduate class where B− is the lowest passing grade) on that part of the test, for repeatedly answering just "true" or "false" with no explanation. Don't let that happen to you.

On each problem, the graders and I mentally start everybody out at −0 (i.e., with a perfect score), and then you accumulate negative points for incorrect answers and/or reasoning, or parts of problems left blank.

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TA (Jacob Fontana). The intent is that the course lecture notes and readings should be sufficient to provide you with all the guidance you need to solve the problems posed below, but you may use other written materials (e.g., the web, journal articles, and books other than those already mentioned in the readings), **provided that you cite your sources thoroughly and accurately**; you will lose (substantial) credit for, e.g., lifting blocks of text directly from `wikipedia` and inserting them into your solutions without full attribution.

If it's clear that (for example) two people have worked together on a part of a problem that's worth 20 points, and each answer would have earned 16 points if it had not arisen from a collaboration, then each person will receive 8 of the 16 points collectively earned (for a total score of 8 out of 20), and I reserve

the right to impose additional penalties at my discretion. If you solve a problem on your own and then share your solution with anyone else, you're just as guilty of illegal collaboration as the person who took your solution from you, and both of you will receive the same penalty. This sort of thing is necessary on behalf of the many people who do not cheat, to ensure that their scores are meaningfully earned. In the AMS graduate class in 2012 mentioned above, five people failed the class because of illegal collaboration; don't let that happen to you.

Those of You who are using `LaTeX` or some other word-processing environment to prepare Your solutions can stick quote blocks below each question, into which You can type Your answers (I suggest that You use **bold** or *italic* font to distinguish Your solutions from the questions). If You're submitting Your answers in longhand, which is perfectly acceptable, You can just write them out on separate sheets of paper, making sure that the grader can easily figure out which chunk of text is the solution to which part of which problem.

Part 1: True/False

[70 total points: 10 points each] For each statement below, say whether it's true or false; if true without further assumptions, briefly explain why it's true; if it's sometimes true, give the extra conditions necessary to make it true; if it's false, briefly explain how to change it so that it's true and/or give an example of why it's false. If the statement consists of two or more sub-statements and two or more of them are false, you need to explicitly address all of the false sub-statements in your answer.

In answering these questions you may find it helpful to consult Gelman et al. (2014) Chapter 11.

(A) If You can figure out how to do IID sampling from the posterior distribution of interest to You, this will often be more Monte-Carlo efficient than MCMC sampling from the same posterior, especially when the number k of unknown quantities is small.

> ***Solution.*** This is **True**. With IID Monte Carlo the autocorrelations in the column of the Monte Carlo data set are all going to be 0, this leads to the Monte Carlo Standard Error values to be smaller. On the other hand though, when k is large and using IID Monte Carlo sampling, it can be difficult to find a Monte Carlo efficient envelope function. □

(B) A (first-order) Markov chain is a particularly simple stochastic process: to simulate where the chain goes next, You only need to know (i) where it is now and (ii) where it was one iteration ago.

> ***Solution.*** This is **False**. This is because as we have learned, to simulate where the chain is going, we only have to know where it is now. So to make it **True**, we just remove the need for the 2nd condition. □

(C) The bootstrap is a frequentist simulation-based computational method (with ties to Bayesian nonparametrics) that can be used to create approximate confidence intervals for population summaries even when the population distribution of the outcome variable y of interest is not known; for example, if (by exchangeability, implied by the problem context) all You know is that Your observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ are IID from *some* distribution with finite mean μ and finite SD σ, You can use

the bootstrap to build an approximate confidence interval for μ even though You don't know what the population distribution is.

> ***Solution.*** This is **True** based on what we covered in lecture on 2/16. We also have seen through the bootstrap example on case study 2, we only needed the sufficient statistics, 403 and 72, to make an approximate confidence interval without having to assume a sampling model! □

(D) In MCMC sampling from a posterior distribution, You have to be really careful to use a monitoring period of just the right length, because if the monitoring goes on for too long the Markov chain may drift out of equilibrium.

> ***Solution.*** This is **False**. This is obviously false because it is a contradiction on the definition of equilibrium. Once equilibrium is reached it stays there indefinitely, meaning it can't drift out of it. □

(E) Simulation-based computational methods are needed in Bayesian data science (inference, prediction and decision-making) because conjugate priors don't always exist and high-dimensional probability distributions are difficult to summarize algebraically.

> ***Solution.*** This is **True**. Since as we've seen in class, when $k > 2$ conjugate priors are pretty rare. We've also seen that it is indeed the case that when k is large, PDF's will become very difficult to work with algebraically. Which is why roughly around 1760-1990 doing Bayesian computation was extremely difficult. □

(F) In MCMC sampling from a posterior distribution, You have to be really careful to use a burn-in period of just the right length, because if the burn-in goes on for too long the Markov chain will have missed its chance to find the equilibrium distribution.

> ***Solution.*** This is **False**. We are able to burn in as long as we'd like and won't lose any information. This follows similarly to what we discussed in THT 2 2B-g. Once we begin the chain at iteration 0, even with a bad starting value, the chain will reach equilibrium and stay in there by definition. Which means it's impossible to burn in for so long that we "miss" the equilibrium period. So what this might want to say is that we should concern ourselves with burn in periods that are too short, since we might not yet have reached equilibrium, and thereby might include draws in our monitoring that are not from the correct equilibrium distribution. □

(G) You're MCMC sampling from a posterior distribution for a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, in which $k \geq 1$ is a finite integer. During the monitoring period, the column in the MCMC data set for a component of θ (θ_j, say) behaves like an autoregressive time series of order 1 ($AR_\rho(1)$) with estimated first-order autocorrelation $\hat{\rho}_j = 0.992$. As usual, You'll use the sample mean $\bar{\theta}_j^*$ of the monitored draws θ_{ij}^* as Your Monte Carlo estimate of the posterior mean of θ_j. To achieve the same estimated Monte Carlo standard error for $\bar{\theta}_j^*$ that You would have been able to attain if You could have done IID sampling, Your MCMC monitoring sample size would have to be about 250 times bigger than the length of the IID monitoring run.

Solution. This is **True**. As we have seen in class the monitoring sample size works out to be,

$$M^*_{MCMC} = M^*_{IID}\left(\frac{1+\hat{\rho}_j}{1-\hat{\rho}_j}\right)$$

and plugging this into a calculator for $\hat{\rho} = 0.992$, we get, 249. Which aligns with the problem statement. □