University of California, Santa Cruz
Department of Applied Mathematics and Statistics
Baskin School of Engineering

AMS 206 (Applied Bayesian Statistics)

Take-Home Test 1 (with typo[s!] corrected)

Target due date **8 Feb 2018**; drop-dead due date **20 Feb 2018**

Here are the ground rules: this test is open-book and open-notes, and consists of two problems (true/false and calculation); **each of the 12 true/false questions is worth 10 points, and the calculation problem is worth 144 total points, for a total of 264 points**.

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 non-extra-credit problems, 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. On extra-credit problems, the usual outcome is that you go forward (in the sense that your overall score goes up) or you at least stay level, but please note that it's also possible to go backwards on such problems (e.g., if you accumulate +3 for part of an extra-credit problem but −4 for the rest of it, for saying or doing something egregiously wrong).

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TA (Daniel Kirsner). 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 paper 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 (because people from your cultural background routinely do this, or out of pity, or kindness, or whatever motive you may believe you have; it doesn't matter), 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.

In class I've demonstrated numerical work in `R`; you can (of course) make the calculations and plots requested in the problems below in any environment you prefer (e.g., `Matlab`, ...).

1 True/False

[120 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 (and — extra credit — what its implications are for statistical inference); 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.

(A) You're about to spin a roulette wheel, which will result in a metal ball landing in one of 38 slots numbered $\Omega = \{0, 00, 1, 2, \ldots, 36\}$; 18 of the numbers from 1 to 36 are colored red, 18 are black, and 0 and 00 are green. You regard this wheel-spinning as fair, by which You mean that all 38 elemental outcomes in Ω are equipossible. Under Your assumption of fairness, the classical (Pascal-Fermat) probability of getting a red number on the next spin exists, is unique, and equals $\frac{18}{38}$.

(B) Under the same conditions as (A), the Kolmogorov (frequentist) probability of getting a red number on the next spin exists, is unique, and equals $\frac{18}{38}$.

(C) Repeat (A) and (B) but removing the assumption that the wheel-spinning is fair, and not replacing it with any other assumption about the nature of the data-generating process (taking the outcomes of the wheel spins as data).

(D) In the Bernoulli sampling model, in which $(Y_1, \ldots, Y_n \,|\, \theta\, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$, the sum $s_n = \sum_{i=1}^{n} y_i$ of the observed data values $\boldsymbol{y} = (y_1, \ldots, y_n)$ is sufficient for inference about θ, and this means that in this model You can throw away the data vector $\boldsymbol{y}$ and focus only on s_n without any loss of information whatsoever.

(E) In learning how to do a good job on the task of uncertainty quantification, it's good to know quite a bit about both the Bayesian and frequentist paradigms, because (a) the Bayesian approach to probability ensures logical internal consistency of Your uncertainty assessments but does not guarantee good calibration, and (b) the frequentist approach to probability provides a natural framework in which to see if Your Bayesian answer *is* well-calibrated.

(F) The $\text{Beta}(\theta \,|\, \alpha, \beta)$ parametric family of distributions is useful as a source of prior distributions when the sampling model is as in (D), because all distributional shapes (symmetric, skewed, multimodal, ...) on $(0, 1)$ are realizable in this family.

(G) Specifying the ingredients $\{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta\, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B})\}$ in Your model for Your uncertainty about an unknown θ (in light of background information $\mathcal{B}$ and data D) is typically easy, because in any given problem there will typically be one and only one way to specify each of these ingredients; an example is the Bernoulli sampling distribution $p(D \,|\, \theta\, \mathcal{B})$ arising uniquely, under exchangeability, from de Finetti's Theorem for binary outcomes.

(H) In trying to construct a good uncertainty assessment of the form $P(A \,|\, \mathcal{B})$, where A is a proposition and $\mathcal{B}$ is a proposition of the form (B_1 and B_2 and $\ldots$ and B_k), You should try hard not to condition on any propositions B_i that are false, because that would be the probabilistic equivalent of dividing by zero.

(I) The kind of objectivity in probability assessment sought by people like Venn, in which all reasonable people would agree on the assessed value, is often impossible to achieve, because all such assessments are conditional on the (1) assumptions, (2) judgments and (3) background information of the person making the probability assessment, and different reasonable people can differ along any of those three dimensions.

(J) When making a decision in the face of uncertainty about an unknown θ, after specifying Your action space $(\mathcal{A} \,|\, \mathcal{B})$ and utility function $U(a, \theta \,|\, \mathcal{B})$ and agreeing on the convention that large utility values are to be preferred over small ones, the optimal decision is found by maximizing $U(a, \theta \,|\, \mathcal{B})$ over all $a \in (\mathcal{A} \,|\, \mathcal{B})$.

(K) One reason that Bayesian inference was not widely used in the early part of the 20th century was that approximating the (potentially high-dimensional) integrals arising from this approach was difficult in an era when computing was slow and the Laplace-approximation technique had been forgotten.

(L) Jaynes (2003, pp. 21–22) makes a useful distinction between {reality} (epistemology) and {Your current information about reality} (ontology); this distinction is useful in probabilistic modeling because {the world} does not necessarily change every time {Your state of knowledge about the world} changes.

2 Calculation

(A) *[48 total points]* Consider the HIV screening example we looked at in class, in which $A =$ {the patient in question is HIV positive}, $+ =$ {ELISA says he's HIV positive} and $- =$ {ELISA says he's HIV negative}. Let p stand for the prevalence of HIV among people similar to this patient (recall that in our example $p = 0.01$), and let ϵ and π stand for the sensitivity and specificity of the ELISA screening test, respectively (in our case study $\epsilon = 0.95$ and $\pi = 0.98$).

(i) By using Bayes's Theorem (in probability or odds form), write down explicit formulas in terms of p, ϵ, and π for the *positive predictive value* (PPV), $P(A \,|\, +)$, and *negative predictive value* (NPV), $P(\text{not } A \,|\, -)$, of screening tests like ELISA (ELISA's PPV and NPV with patients like the one in our case study were 0.32 and 0.99948, respectively). These formulas permit analytic study of the tradeoff between PPV and NPV. *[10 points]*

(ii) Interest focused in class on why ELISA's PPV is so bad for people, like the guy we considered in the case study, for whom HIV is relatively rare ($p = 0.01$).

(a) Holding ϵ and π constant at ELISA's values of 0.95 and 0.98, respectively, obtain expressions (from those in 2(A)(i)) for the PPV and NPV as a function of p, and plot these functions as p goes from 0 to 0.1. *[8 points]*

(b) Show by means of Taylor series that in this range the NPV is closely approximated by the simple linear function $(1 - 0.056\,p)$. *[6 points]*

(c) How large would p have to be for ELISA's PPV to exceed 0.5? 0.75? *[4 points]*

(d) What would ELISA's NPV be for those values of p? *[4 points]*

(e) Looking at both PPV and NPV, would You regard ELISA as a good screening test for subpopulations with (say) $p = 0.1$? Explain briefly. *[4 points]*

(iii) Suppose now that p is held constant at 0.01 and we're trying to improve ELISA for use on people with that prevalence of HIV, where "improve" for the sake of this part of the problem means raising the PPV while not suffering too much of a decrease (if any) of the NPV. ELISA is based on the level L of a particular antibody in the blood, and uses a rule of the form {if $L \geq c$, announce that the person is HIV positive}. This means that if You change c the sensitivity and specificity change in a tug-of-war fashion: altering c to make ϵ go up makes π go down, and vice versa.

(a) By using the formulas in 2(A)(i), show that as ϵ approaches 1 with π no larger than 0.98, the NPV will approach 1 but the biggest You can make the PPV is about 0.336. Thus if we want to raise the PPV we would be better off trying to increase π than ϵ. *[4 points]*

Suppose there were a way to change c that would cause π to go up while holding ϵ arbitrarily close to 0.95.

(b) Show that π would have to climb to about 0.997 to get the PPV up to 0.75. *[4 points]*

(c) Is the NPV still at acceptable levels under these conditions? Explain briefly. *[4 points]*

(B) *[96 total points]* (Bayesian conjugate inference with the Exponential distribution) In a consulting project that one of my Ph.D. students and I worked on at the University of Bath in England before I came to Santa Cruz, a researcher from the Department of Electronic and Electrical Engineering (EEE) at Bath wanted help in analyzing some data on failure times for a particular kind of metal wire (in this problem, failure time was defined to be the number of times the wire could be mechanically stressed by a machine at a given point along the metal before it broke). The $n = 14$ raw data values y_i in one part of her experiment, arranged in ascending order, were

495 541 1461 1555 1603 2201 2750 3468 3516 4319 6622 7728 13159 21194

From the context $\mathbb{C}$ of this problem, Your uncertainty about these data values before they were observed is exchangeable, which implies that it's appropriate to model the y_i as conditionally IID, but from what distribution? The simplest model for failure time data involves the *Exponential* distribution:

$$(y_i \,|\, \lambda \, \boldsymbol{E} \, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Exponential}(\lambda): \quad \text{i.e.,} \quad p(y_i \,|\, \lambda \, \boldsymbol{E} \, \mathcal{B}) = \left\{ \begin{array}{cc} \frac{1}{\lambda} \exp(-\frac{y_i}{\lambda}) & y_i > 0 \\ 0 & \text{otherwise} \end{array} \right\} \quad (1)$$

for some $\lambda > 0$, in which $\boldsymbol{E}$ stands for the Exponential sampling distribution assumption (which is not part of $\mathcal{B}$, since it's not implied by problem context but has instead been chosen for simplicity). (**NB** This distribution can be parameterized either in terms of λ or $\frac{1}{\lambda}$; whenever it comes up, You need to be careful which parameterization is in use.)

(i) To see if this model fits the data set given above, You can make an *Exponential probability plot*, analogous to a Gaussian quantile-quantile plot (*qqplot*) to check for Normality.

In fact the idea works for more or less any distribution: You plot

$$y_{(i)} \quad \text{(vertical axis)} \quad \text{versus} \quad F^{-1}\left(\frac{i-0.5}{n}\right), \tag{2}$$

where $y_{(i)}$ are the y values sorted from smallest to largest and F is the CDF of the distribution You're considering (the 0.5 is in the numerator to avoid problems at the edges of the data). In so doing You're graphing the data values against an approximation of *what You would have expected for the data values if the CDF of the y_i really had been F*, so the plot should resemble the 45° line if the fit is good.

(a) Work out the CDF $F_Y(y \,|\, \lambda)$ of the Exponential(λ) distribution (parameterized as in equation (1) above) and show that its inverse CDF is given by

$$F_Y(y \,|\, \lambda) = p \iff y = F^{-1}(p \,|\, \lambda) = -\lambda \log(1-p)\,. \tag{3}$$

[8 points]

(b) To use equation (3) to make the plot, we need a decent estimate of λ. Write down the likelihood and log-likelihood functions in this model, simplified as much as You can, and plot them (on different graphs, and with λ ranging on the horizontal scale from 2,000 to 15,000) using the data values given above. Briefly explain why the form of Your log-likelihood function implies that $\bar{y}$, the sample mean, is sufficient for λ in the Exponential sampling model. Show that the maximum likelihood estimate of λ in this model is $\hat{\lambda}_{\text{MLE}} = \bar{y}$, and use this (i.e., take $\lambda = \hat{\lambda}_{\text{MLE}}$ in (3)) to make an Exponential probability plot of the 14 data values above, superimposing the 45° line on it. Informally, does the Exponential model appear to provide a good fit to the data? Explain briefly. *[12 points]*

(ii) By regarding Your likelihood in 2(B)(i)(b) as an unnormalized probability density function for λ, show that the conjugate family for the Exponential(λ) likelihood (parameterized as in (1)) is the set of *Inverse Gamma* distributions $\Gamma^{-1}(\alpha, \beta)$ for $\alpha > 0, \beta > 0$ (**NB** $W \sim \Gamma^{-1}(\alpha, \beta)$ just means that $\frac{1}{W} \sim \Gamma(\alpha, \beta)$; see Table A.1 in Appendix A in Gelman et al. (2014)):

$$\lambda \sim \Gamma^{-1}(\alpha, \beta) \quad \iff \quad p(\lambda) = \left\{ \begin{array}{cc} \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} \exp\left(-\frac{\beta}{\lambda}\right) & \text{for } \lambda > 0 \\ 0 & \text{otherwise} \end{array} \right\}. \tag{4}$$

[8 points]

(iii) By directly using Bayes' Theorem (and ignoring constants), show that the prior-to-posterior updating rule in this model is

$$\left\{ \begin{array}{ccc} (\lambda \,|\, \boldsymbol{IG}) & \sim & \Gamma^{-1}(\alpha, \beta) \\ (Y_i \,|\, \lambda\, \boldsymbol{E}\, \mathcal{B}) & \overset{\text{IID}}{\sim} & \text{Exponential}(\lambda) \end{array} \right\} \Longrightarrow (\lambda \,|\, \boldsymbol{y}\, \boldsymbol{IG}\, \boldsymbol{E}\, \mathcal{B}) \sim \Gamma^{-1}(\alpha + n, \beta + n\bar{y})\,, \tag{5}$$

in which ***IG*** stands for the Inverse Gamma sampling distribution assumption. *[8 points]*

(iv) It turns out that the mean and variance of the $\Gamma^{-1}(\alpha, \beta)$ distribution are $\frac{\beta}{\alpha-1}$ (when $\alpha > 1$) and $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ (as long as $\alpha > 2$), respectively. Use this to write down an explicit formula showing that the posterior mean is a weighted average of the prior and sample means, and conclude from this formula that $n_0 = (\alpha - 1)$ is the prior effective sample size. Note also from the formula for the likelihood in this problem that, when thought of as a distribution in λ, it's equivalent to a constant times the $\Gamma^{-1}(n-1, n\,\bar{y})$ [**typo corrected here**] distribution. *[8 points]*

(v) The researcher from EEE has prior information from another experiment she judges to be comparable to this one: from this other experiment the prior for λ should have a mean of about $\mu_0 = 4{,}500$ and an SD of about $\sigma_0 = 1{,}800$.

(a) Show that this corresponds to a $\Gamma^{-1}(\alpha_0, \beta_0)$ prior with $(\alpha_0, \beta_0) = (8.25, 32625)$, and therefore to a prior sample size of about 7. Is this amount of prior information small, medium or large in the context of her data set? Explain briefly. *[8 points]*

(b) Thinking of each of the prior, likelihood and posterior densities as Inverse Gamma distributions, work out the SDs of each of these information sources, and numerically summarize the updating from prior to posterior by completing this table (show Your work): [**typo corrected in table**]

	λ		
	Prior	Likelihood	Posterior
Mean	4,500		4,858
SD		1,774	

[8 points] Extra credit [xx points]: How do the prior and likelihood SDs combine, at least approximately, to yield the posterior SD? Explain briefly.

(c) Make a plot of the prior, likelihood and posterior distributions on the same graph (with λ ranging on the horizontal scale from 1,000 to 12,000), identifying which curve corresponds to which density (You can use the `R` code on the course web page for the Inverse Gamma density function, or You can write Your own code to evaluate the density in equation (4)). In what sense, if any, is the posterior a compromise between the prior and likelihood? Explain briefly. *[14 points]*

(d) Compute the observed Fisher information with this data set, and use this to compute an estimated standard error for the MLE and construct an approximate 95% frequentist confidence interval for λ. Use the `qgamma` function in `R` (or some other numerical integration routine of Your choice) to work out the left and right endpoints of the 95% central posterior interval for λ (*Hint:* remember the **NB** in 2(B)(ii)), and compare with the frequentist interval. Give two reasons why they're so different in this problem. Is one of them "right" and the other one "wrong'," or are they trying to summarize different amounts and types of information, or what? Explain briefly. *[24 points]*