University of California, Santa Cruz
Department of Statistics
Baskin School of Engineering
Winter 2022

STAT 206 (Applied Bayesian Statistics)

Take-Home Test 1

(See updates in class and by email and `Discord` for the **final deadline**.)

Here are the (process) 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 230 total points, for a total of 350 points**.

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 non-extra-credit problems, 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. 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 (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.

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`, ...). To avoid plagiarism, if you end up using any of the code I post on the course web page or generate during office hours, at the beginning of your Appendix (see below) you can say something like the following:

> *I used some of Prof. Draper's `R` code in this assignment, adapting it as needed.*

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.

Please collect {all of the code you used in answering the questions below} into an Appendix at the end of your document, so that (if you do something wrong) the grader can more accurately give you part credit.

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 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 $(Y_i \,|\, \theta\, \mathcal{B}) \stackrel{\text{IID}}{\sim}$ Bernoulli(θ) for $i = 1, \ldots, n$ (here n is a finite positive integer and $0 < \theta < 1$), 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. (Note that, as usual, we're implicitly conditioning on the known value of n.)

(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 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 by single members of 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 and middle parts 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) *[100 total points]* Consider the HIV screening example we looked at in Case Study (CS) 1, in which $(\theta = 1)$ = (the patient is HIV positive) and $(y_1 = 1)$ = (the blood test says the patient is HIV positive), but now let's make two changes: the time is now 1985, when the

Table 1: *The basic disease screening* (2×2) *table on the probability scale, with* $\theta = (1$ *if the disease is truly present, 0 otherwise),* $y_1 = (1$ *if the screening test says the disease is present, 0 otherwise), and* (α, β, γ) *= (prevalence, sensitivity, specificity).*

		Truth		
		HIV $\oplus$ $(\theta = 1)$	HIV $\ominus$ $(\theta = 0)$	Total
Blood	$\oplus$ $(y_1 = 1)$	TP: $\alpha\,\beta$	FP: $(1-\alpha)\,(1-\gamma)$	$\alpha\,\beta + (1-\alpha)\,(1-\gamma)$
Test	$\ominus$ $(y_1 = 0)$	FN: $\alpha\,(1-\beta)$	TN: $(1-\alpha)\,\gamma$	$\alpha\,(1-\beta) + (1-\alpha)\,\gamma$
	Total	α	$(1-\alpha)$	1

first *enzyme-linked immunosorbent assay (ELISA)* blood test was approved in the U.S. for use in detecting HIV, and You now work for the Red Cross (RC), which maintains a blood bank (from which units of blood for surgeries in hospitals are drawn) and which is extremely interested in not letting HIV into their blood supply. Continuing to use CS 1 notation, let $\alpha = P(\theta = 1 \,|\, \mathcal{B})$ be the prevalence of HIV in people whose background risk factors are summarized in $\mathcal{B}$; and let $\beta = P(y_1 = 1 \,|\, \theta = 1, \mathcal{B})$ and $\gamma = P(y_1 = 0 \,|\, \theta = 0, \mathcal{B})$ be the sensitivity and specificity, respectively, of the first *ELISA* test (let's call it E_1). According to Chappel, Wilson and Dax (2009, *Future Microbiology*, **8**, 963–982), $(\beta, \gamma) = (0.99, 0.95)$ for E_1, so the first test had decent sensitivity but did not reach the same performance level in specificity. Poking around on `www.census.gov`, You'll find that the population of the United States in 1985 consisted of about 238 million people, of whom about $N = 175$ million people were 18 years old or older; let's assume that HIV is concentrated entirely in the 18+ subpopulation (which is true to a good approximation).

The basic (2×2) table for disease screening, in the notation of this problem, is given in Table 1. **Warning:** many published sources use the rows-and-columns convention in Table 1, but some reverse this, with rows for truth and columns for what the screening test says; an example of the latter convention is at

`https://en.wikipedia.org/wiki/Sensitivity_and_specificity` .

The definitions of the four probabilities in the body of the table are as follows: $(TP, FP, FN, TN) = \{$true positive (upper left cell), false positive (upper right), false negative (lower left), true negative (lower right)$\}$.

(i) Where did the four entries in the body of Table 1 (not the margins) come from? As an example, by making an easy calculation, briefly explain why the upper-left entry is $\alpha\,\beta$; why the same logic applies to the lower-right entry; and how the other entries are then immediately calculated. *[15 points]*

(ii) Use this table to write down explicit formulas in terms of (α, β, γ) for two frequently-used quantities in disease screening that we haven't looked at yet: the *positive predictive value* (PPV, also known as the *precision*), $P(\theta = 1 \,|\, y_1 = 1)$, and the *negative predictive value* (NPV, with a similar interpretation for negative test results), $P(\theta = 0 \,|\, y_1 = 0)$, of screening tests such as E_1. How do the PPV and NPV relate to the *false discovery* and *false omission rates* (FDR $= \frac{FP}{FP+TP}$, FOR $= \frac{FN}{FN+TN}$)? Explain briefly. *[20 points]*

(iii) The Centers for Disease Control and Prevention (CDC, not CDCP, for some reason) estimated in 2016 that the U.S. prevalence of HIV in 1985 was based on about 500,000

Table 2: *Partially-filled-out table of expected numbers of people receiving HIV diagnoses under the Congressperson's plan.*

		Truth		
		HIV $\oplus$ ($\theta = 1$)	HIV $\ominus$ ($\theta = 0$)	Total
Blood Test	$\oplus$ ($y_1 = 1$)	495,000	$N\,(1-\alpha)\,(1-\gamma)$	$N\,[\alpha\,\beta + (1-\alpha)\,(1-\gamma)]$
	$\ominus$ ($y_1 = 0$)	$N\,\alpha\,(1-\beta)$	$N\,(1-\alpha)\,\gamma$	165,780,000
	Total	$N\,\alpha$	174,500,000	N

cases, for a prevalence rate in the 18+ subpopulation of $\frac{500000}{175000000} = \alpha^* \doteq 0.00286$, about 0.3% (roughly the same as the U.S. prevalence rate today). The Red Cross (RC) would not have been privy to this information in 1985, but assuming that HIV status and the blood-donation choice mechanism are independent, which is almost certainly upper-bounding for α, would give α^* as the RC prevalence. Use this value for α and the (β, γ) values for E_1 to compute the PPV, NPV, FPR and FNR values defining the blood-screening real-world environment facing the RC in 1985. Would You say that E_1 was highly successful at keeping HIV out of the RC blood supply in 1985? Explain briefly. *[25 points]*

(iv) Holding (β, γ) at the E_1 values and varying α from 0 to (say) $10\,\alpha^*$, plot the PPV and NPV as functions of α. How sensitive were each of these quantities to prevalence in the 1985 RC environment? Explain briefly. *[15 points]*

Shortly after E_1 was approved in 1985, a member of the U.S. Congress made a speech on the floor of the House of Representatives expressing the opinion that HIV was such a serious public health threat that everyone 18+ years old should be tested with E_1. The goal in this final part of the problem is to fill out a new version of Table 1 with numbers quantifying what would have happened to the $N = 175$ million Americans under this Congressperson's plan. If we knew for sure that $\alpha = \alpha^*$, we could just use that value of α and the already-established values of (β, γ), and multiply all of the resulting entries in Table 1 by N, but we don't know that for sure. Consider α an unknown quantity (in STAT 131 we would have called it a random variable) with expected value $E(\alpha \,|\, \mathcal{B}) = \alpha^*$.

(v) By looking at the form of all 9 of the entries in Table 1 (including the margins) as functions of α (and remembering basic properties of expectation from STAT 131), briefly explain why we can obtain a table of *expected* cell and margin counts just by multiplying all of the entries in Table 1 by N and then substituting in $(\alpha, \beta, \gamma) = \left(\frac{500000}{175000000}, 0.99, 0.95\right)$. Complete Table 2 by filling in the empty cells and margins; I've given You a headstart on some of them. Briefly summarize the likely good and bad outcomes of the Congressperson's plan, when viewed as an instance of national health policy. (*Hint:* The first Western Blot test for HIV, which was more accurate than E_1, was not developed until 1987.) In Your view, would the good outcomes outweigh the bad, or the other way around, or is it hard to come to a clear judgment? Explain briefly. (Note that we're not doing a complete *cost-benefit* analysis here, since we've not taken into account how much administering 175,000,000 E_1 tests would cost in time and money.) *[25 points]*

(B) *[130 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

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

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

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

[10 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 $\lambda = \hat{\lambda}$ and $p = \left(\frac{i-0.5}{n}\right)$ in equations (2) and (3)) to make an Exponential probability plot of the 14 data values above (i.e., plot the sorted y values on the vertical axis against $F^{-1}\left(\frac{i-0.5}{n} \,|\, \hat{\lambda}_{\text{MLE}}\, \boldsymbol{E}\right)$, superimposing the 45° line on it. Informally, does the Exponential model appear to provide a good fit to the data? Explain briefly. *[35 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)):

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

in which $\boldsymbol{IG}$ stands for the Inverse Gamma sampling distribution assumption *[5 points]*.

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

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

where $\boldsymbol{y} = (y_1, \ldots, y_n)$. *[10 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 $\Gamma^{-1}(n-1, n\,\bar{y})$ distribution. *[10 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. *[10 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 Table 3 (show Your work) *[10 points]*.

(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. *[15 points]*

Table 3: *Bayesian updating in the wire-failure case study.*

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

(d) Compute the observed information with this data set, and use this to compute an estimated standard error for the MLE and construct an approximate 99.9% 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 99.9% 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. *[25 points]*