Prof. David Draper
University of California, Santa Cruz
Department of Statistics
Baskin School of Engineering
Winter 2024

Name: **Devanathan Nallur Gandamani (2086936)**

STAT 206 (Applied Bayesian Statistics)

Take-Home Test 1 ***(340 Total Points, Plus 140 Possible Extra Credit Points, For a Total of 480 Points)***

(Please see updates in class — by email, in `Discord`, and in `Canvas Announcements` — 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 required calculation problem is worth 230 total points, and the extra credit calculation problem is worth 140 points, for a total of 480 possible points**. You don't need to complete all of the extra credit problem to get extra points; any part of the problem that you complete correctly will add numerator points and no denominator points to your course score

$$\frac{\text{total correct points across all assignments}}{\text{total possible non-extra-credit points across all assignments}} . \tag{1}$$

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 TAs. 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.

> ***Under UCSC policies, submission of your solutions constitutes acceptance of the ethical rules stated above.***

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., `python`, `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 graders can more accurately give you part credit.***

In what follows I've not left blank spaces for your solutions; `LaTeX` (and other technical text processing) people can insert quote blocks below each question, in bold or bold-italic font and perhaps also in a non-black legible color; people submitting handwritten solutions can use extra sheets of paper, as long as each solution fragment is clearly marked with the question it's answering.

(I) True/False

[110 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.

(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 $\frac{18}{38}$.

Solution:

True

- Classical Probability can be calculated using below formula :

$$P(E) = \frac{\text{Number of outcomes favorable to } E}{\text{Total number of equally likely outcomes}}$$

  P(E) - Probability of getting a red number

- According to question, Number of outcomes favorable to E = 18 Total number of equally likely outcomes = 36
- Therefore,

$$P(E) = \frac{18}{36}$$

- According to classical interpretation of probability, " It assumes that all outcomes in the sample space are equally likely to occur and that the sample space is finite."
- Since Roulette wheel is fair, and we have finite sample space of total 38 slots.Therefore, P(E) which is frequency of landing on a red number (18) compared to all possibel outcomes (38) is 18/36]

(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 $\frac{18}{38}$.

True

- According to frequentist interpretation of probability, "the probability of an event is determined by its long-run frequency of occurrence under identical conditions."
- Since Roulette wheel is fair, [each of the 38 possible outcomes (slots on the wheel) has an equal chance of occurring on any given spin and that the process is repeated many times. Therefore, This fairness assumption coincides with the frequentist approach by calculating that, in an infinitely long series of spins, the relative frequency of landing on a red number (there are 18 of them) compared to all potential outcomes (38 in total) will approach 18/36]

(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). False

- Explanation :

For (A)

- **Normal Scenario:** We say that each number has an equal chance of coming up. So, if there are 18 reds out of 38 total numbers, the chances of receiving a red are $\frac{18}{38}$.
- **Without Fairness:** If the wheel is not fair, we can no longer claim that each number has an equal probability. Specific numbers may appear more frequently because the wheel is old, imbalanced, or has a manufacturing problem. However, we need to understand how the wheel is skewed to use the traditional probability method to calculate the probability of getting a red.

For (B)

- **Normal Scenario:** This method examines what happens when you spin the wheel repeatedly. If everything is fair, spinning the wheel repeatedly should result in reds appearing 18 out of 38 times on average.
- **Without Fairness:** Even if the wheel isn't fair, we can see what happens after several spins. If red appears more or less frequently than 18 out of 38 times after several spins, we utilize this new figure to calculate the probability of red occurring. This way, we learn from what's happening with the wheel.

(D) You're observing a binary data stream, and You have sampling model $[SM]$ uncertainty because You're not sure about the IID part of the usual Bernoulli assumptions, in which $(Y_i \,|\, [SM \colon \underline{\text{Ber}}]\, \theta\, \mathcal{B}) \overset{\text{IID}}{\sim}$ Bernoulli(θ) for $i = 1, \ldots, n$ (here n is a finite positive integer, $0 < \theta < 1$ is the unknown mean of the data stream, and Ber stands for the Bernoulli sampling model). In this situation the vector (n, s) is sufficient for inference about θ, where $s = \sum_{i=1}^{n} y_i$ is the sum of the observed data values $\boldsymbol{y} = (y_1, \ldots, y_n)$, and this means that You can throw away the data vector $\boldsymbol{y}$ and focus only on s without any loss of relevant information whatsoever.

True

The statement is True considering the following:

1. The statement refers to a binary data stream evaluated using a Bernoulli sampling technique. Each observation has only two possibilities: success (1) or failure (0).

The IID assumption is important because it indicates that each observation is selected independently from the same Bernoulli distribution with success probability θ, where θ lies between 0 and 1.

2. The vector (n, s) is sufficient for inference about θ,

n - number of trials

s - sum of successes.

Sufficiency means (n, s) includes all the necessary knowledge to estimate θ, making the particular outcomes in y redundant for this specific inferential purpose. This statistical concept ensures that information about θ is not lost while reducing data from the entire vector y to summary statistics (n, s).

This highlights a fundamental statistical principle: condensing data into sufficient statistics can speed up processing while maintaining inference quality.

(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.

True

(a) The Bayesian approach to probability offers a consistent framework for updating beliefs in the presence of data. It incorporates prior knowledge and uses Bayes' theorem to update beliefs. Bayesian approaches maintain logical coherence in uncertainty assessments but do not guarantee that these evaluations adequately reflect the underlying uncertainty in the data generation process. Calibration difficulties can develop if the prior distribution or likelihood function used in Bayesian analysis is incorrectly stated or the model assumptions do not accurately reflect the genuine underlying process.

(b) The frequentist approach to probability focuses on the long-term behavior of random processes and using sample statistics to draw conclusions about population parameters. It establishes a rigorous statistical inference framework, especially when repeated sampling or observation is possible. It provides objective methods for drawing conclusions about unknown parameters and evaluating hypotheses using observable data.

Finally, comprehending the Bayesian and frequentist paradigms offers complementary viewpoints. The Bayesian approach stresses logical coherence in uncertainty assessments, whereas the frequentist approach frequently focuses on qualities such as frequentist coverage and estimator unbiasedness. Combining insights from both perspectives can result in a more complete understanding of uncertainty and statistical inference.

(F) The Beta$(\theta \,|\, \alpha, \beta)$ parametric family of distributions is useful as a prior model $[PM]$ when the sampling model $[SM]$ is as in (D), because all distributional shapes (symmetric, skewed, multimodal, ...) on $(0, 1)$ are realizable by single members of this family.

False

Explanation:

The Beta distribution is limited to the interval (0, 1) and has two shape parameters, α and β. While it is adaptable enough to describe a wide range of distributions on the unit interval, it does not cover all potential shapes (symmetric, skewed, multimodal, etc.). For example, a bimodal distribution with one peak near 0 and another near 1 cannot be accurately represented by a single Beta distribution.

To make this statement TRUE, it must be revised to reflect the Beta distribution family's restrictions or to clarify the situations under which it may accurately represent the intended distributional shapes.

(G) Specifying the ingredients $\{p(\theta \,|\, [PM]\, \mathcal{B}), p(D \,|\, [SM]\, \theta\, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B})\}$ 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 $p(D \,|\, [SM:\,]\, \theta\, \mathcal{B})$ arising uniquely, under exchangeability, from de Finetti's Theorem for binary outcomes.

True

The statement is correct because describing the ingredients $\{p(\theta|PM), p(D|SM, \theta), p(A|B), U(a, \theta|B)\}$ in a model for uncertainty about an unknown parameter θ, given background information B and observed data D, is often straightforward.

Each ingredient has a specific function:

- $p(\theta|PM)$ - represents the prior distribution for θ
- $p(D|SM, \theta)$ - represents the likelihood function
- $p(A|B)$ - represents the probability of a proposition A given background information B
- $U(a, \theta|B)$ - represents any additional uncertainty or randomness in the model.

In many circumstances, these ingredients can be specifically determined depending on the problem context and accessible information.

For example, in situations where we're dealing with binary outcomes (like flipping a coin), the likelihood function can be uniquely determined, making it simpler to build the model.

(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_b) (in which b is a finite positive integer), You should try hard not to condition on any propositions B_i that are false, because that would be the probabilistic equivalent of trying to quantify $\frac{0}{0}$.

True

When developing an uncertainty assessment, it is critical to avoid conditioning on incorrect assumptions to ensure validity. Conditioning on false assertions can result in logical errors and erroneous probability assignments since it effectively assigns a probability of zero to that branch of the sample space, which violates probability principles. This is similar to attempting to estimate the chance of an impossible, fundamentally useless event.

For example, suppose the assertion "It is snowing" is known to be false. In that case, conditioning on it might result in a nonsensical probability assignment for rain-dependent events, such as outdoor activities.

(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.

True

The statement is valid since obtaining total objectivity in probability analysis can often be impossible. This is because the probability evaluation is based on the assessor's assumptions, judgments, and prior information. People may have different beliefs, judgments, and background information, resulting in variable probability evaluations.

For example, let's analyze the probability of rain tomorrow. One individual may evaluate the chance using meteorological data, while another may use past weather trends or intuition, resulting in differing conclusions.

(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 $a \in (\mathcal{A} \,|\, \mathcal{B})$.

True

This statement is correct. When faced with uncertainty about an unknown parameter θ and given the action space $(A|B)$ and utility function $U(a, \theta|B)$, which specifies the implications of alternative actions under uncertainty, the optimal solution is to maximize the utility function over all feasible actions $a \in (A|B)$. This involves selecting the activity that maximizes the expected value while considering all possible outcomes and their probabilities.

For example, when determining whether to invest in a stock, one might weigh the prospective returns and risks associated with several investment techniques before selecting the one that maximizes expected utility.

(K) 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.

True

This statement is correct. Jaynes' distinction between actuality (epistemology) and present knowledge of reality (ontology) is helpful in probabilistic modeling. This difference recognizes that humans' understanding of the universe (ontology) differs from fundamental reality (epistemology). In probabilistic modeling, our knowledge of the world can shift without the world changing.

For example, assume a person says there is a 50 percent chance of rain tomorrow based on the weather prediction. Suppose he later obtains more accurate meteorological data suggesting a more significant possibility of rain. In that case, his knowledge (ontology) changes, but the underlying reality (whether it will rain or not) remains constant until tomorrow.

(II) Calculation

(A) ***[230 total points]*** (Likelihood and 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 was exchangeable, which implies that it's appropriate to model the y_i as conditionally IID, but from what distribution?

The simplest sampling model $[SM]$ for failure time data involves the *Exponential* distribution:

$$\begin{aligned} (Y_i \,|\, [SM\!:\!\mathbb{E}]\, \lambda\, \mathcal{B}) &\overset{\text{IID}}{\sim} \text{Exponential}(\lambda)\,, \quad \text{i.e.,} \\ p(y_i \,|\, [SM\!:\!\mathbb{E}]\, \lambda\, \mathcal{B}) &= \frac{1}{\lambda} \exp(-\frac{y_i}{\lambda})\, I(y_i > 0)\, I(\lambda > 0)\,, \end{aligned} \tag{2}$$

in which (*) $I(A)$ is 1 if the proposition A is true and 0 otherwise and (**) $\mathbb{E}$ 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.)

(1) To see if this model fits the data set given above, when adopting the ***cheating approach to*** $[SM]$ ***specification*** (which we will in this problem; more satisfying options for dealing with $[SM]$ uncertainty will be covered later), 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 on $\mathbb{R}$: You plot

$$y_{(i)} \quad \text{(vertical axis)} \quad \text{versus} \quad F_Y^{-1}\left(\frac{i-0.5}{n}\right), \tag{3}$$

where $y_{(i)}$ are the y values sorted from smallest to largest and F_Y is the CDF of the random variable Y that 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_Y**, so the plot should resemble the 45° line if the fit is good (this is an example of the **Probability Integral Transform** idea from STAT 131).

(a) Work out the CDF $F_Y(y \,|\, [SM\!:\!\mathbb{E}]\, \lambda)$ of the Exponential(λ) distribution (parameterized as in equation (2) above) and show that its inverse CDF is given by

$$F_Y(y \,|\, [SM\!:\!\mathbb{E}]\, \lambda) = p \iff y = F_Y^{-1}(p \,|\, [SM\!:\!\mathbb{E}]\, \lambda) = -\lambda \, \log(1-p)\,. \tag{4}$$

[10 points]

(b) To use equation (4) to make the plot, we need a decent estimate of λ.

(i) 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; include Your plot in Your solutions. ***[20 points]***

(ii) Briefly explain why the form of Your log-likelihood function implies that $\bar{y}$, the sample mean, is sufficient for λ in the Exponential $[SM]$ (along with n, of course). ***[10 points]***

(iii) Show that the maximum likelihood estimate of λ in this model is $\hat{\lambda}_{\text{MLE}} = \bar{y}$. ***[10 points]***

(iv) Use (iii) (i.e., take $\lambda = \hat{\lambda}_{\text{MLE}}$ and $p = \left(\frac{i-0.5}{n} \right)$ in equations (3) and (4)) to make an Exponential probability plot of the 14 data values above (i.e., plot the sorted y values on the vertical axis against $F_Y^{-1}\left(\frac{i-0.5}{n} \,|\, [SM\!:\!\mathbb{E}]\, \hat{\lambda}_{\text{MLE}} \right)$, superimposing the 45° line on it; include Your plot in Your solutions. ***[10 points]***

(v) Informally, does the Exponential sampling model appear to provide a good fit to the data? Explain briefly. ***[10 points]***

(2) By regarding Your likelihood in (II)(A)(1)(b)(i) as an unnormalized probability density function for λ, show that the conjugate family for the Exponential(λ) likelihood (parameterized as in (2)) 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 $\frac{1}{W}$ follows the Gamma distribution $\Gamma(\alpha, \beta)$; see Table A.1 in Appendix A in Gelman et al. (2014)):

$$\lambda \sim \Gamma^{-1}(\alpha, \beta) \quad \Longleftrightarrow \quad p(\lambda \,|\, \boldsymbol{\Gamma^{-1}}) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{-(\alpha+1)} \exp\left(-\frac{\beta}{\lambda} \right) I(\lambda > 0)\,, \tag{5}$$

in which $\boldsymbol{\Gamma^{-1}}$ stands for the Inverse Gamma distributional assumption ***[10 points]***.

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

$$\left\{ \begin{array}{ccc} (\lambda \,|\, [PM\!:\!\boldsymbol{\Gamma^{-1}}]) & \sim & \Gamma^{-1}(\alpha, \beta) \\ (Y_i \,|\, [SM\!:\!\mathbb{E}]\, \lambda\, \mathcal{B}) & \stackrel{\text{IID}}{\sim} & \text{Exponential}(\lambda) \end{array} \right\} \Longrightarrow$$

$$(\lambda \,|\, \boldsymbol{y}\, [PM\!:\!\boldsymbol{\Gamma^{-1}}]\, [SM\!:\!\mathbb{E}]\, \mathcal{B}) \sim \Gamma^{-1}(\alpha + n, \beta + n\bar{y})\,, \tag{6}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$. ***[10 points]***

(4) 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.

(a) 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. ***[10 points]***

(b) Note also from the formula for the likelihood in this problem that, when thought of as a distribution for λ, it's equivalent to a constant times the $\Gamma^{-1}(n - 1, n\,\bar{y})$ distribution. ***[10 points]***

(5) The researcher from EEE has prior information from another experiment that 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) Quantifying the amount of prior information:

(i) 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. ***[10 points]***

(ii) Is the amount of prior information in (i) 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 1 (show Your work) ***[20 points]***.

Table 1: *Bayesian updating in the wire-failure case study.*

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

(c) Visualizing the Bayesian inferential story:

(i) 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, (**) download the Inverse Gamma package from `CRAN`, or (***) write Your own code to evaluate the density in equation (5)); include Your plot in Your solutions. ***[10 points]***

(ii) In what sense, if any, is the posterior a compromise between the prior and likelihood? Explain briefly. ***[10 points]***

(d) Let's conclude this case study by making an inferential summary for λ.

(i) 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. ***[20 points]***

(ii) Use the `qinvgamma` function (from `CRAN`) 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, and compare with the frequentist interval. ***[20 points]***

(iii) Give two reasons why the likelihood and Bayesian intervals are 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. ***[20 points]***

(B) ***[140 total points of extra credit]*** Study the documents called

```
stat-206-lecture-notes-part-x.pdf
```

for `x` from 1 to 4 on the `Pages` tab of the course `Canvas` web pages as preparation for this problem.

Consider the HIV screening example described in what's called Case Study (CS) 1 in the lecture notes files mentioned above, in which $(\theta = 1) =$ (the patient is HIV positive) and $(y_1 = 1) =$ (the blood test says the patient is HIV positive), but let's make two changes: the time is now 1985, when the 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*

Table 2: *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 Test	$\oplus$ $(y_1 = 1)$	$TP\!: \ \alpha \, \beta$	$FP\!: \ (1 - \alpha) \, (1 - \gamma)$	$\alpha \, \beta + (1 - \alpha) \, (1 - \gamma)$
	$\ominus$ $(y_1 = 0)$	$FN\!: \ \alpha \, (1 - \beta)$	$TN\!: \ (1 - \alpha) \, \gamma$	$\alpha \, (1 - \beta) + (1 - \alpha) \, \gamma$
	Total	α	$(1 - \alpha)$	1

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 2. **Warning:** Many published sources use the rows-and-columns convention in Table 2, but some reverse this, with rows for truth and columns for what the screening test says; an example of the latter convention is at `this Wikipedia` page.

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)}.

(1) ***[30 total points for this part of this problem]*** Where did the four entries in the body of Table 2 (not the margins) come from? As an example, by making easy probability calculations, briefly explain

(a) why the upper-left entry is $\alpha \, \beta$, ***[10 points]***

(b) why the same logic applies to all of the other entries ***[10 points]*** , and

(c) how the row and column margin totals may then be calculated. ***[10 points]***

Solution:

a): True positives are instances where both the test indicates HIV positivity and the subject indeed has HIV. The problem specifies that α represents the prevalence of HIV among those with risk factors of β. Additionally, it states that β = P(y1 = 1 $|\theta$ = 1) represents the sensitivity of our test. Therefore, true positives occur when the test accurately predicts HIV, which is β multiplied by the prevalence of HIV overall, (1 - α) (1 - β). Hence, our true positives can be calculated as α β.

b): False Positives : When no HIV but test is positive. (1 - α) is probability that someone doesn't have HIV in general, multiplying that by (1 - γ) which is the rate of how often our test will give a false positive given the patient is well, thus we multiplying how often someone doesn't have HIV in general by how often the test returns a positive if the person is well (1 - α)(1 - β) gives us the proportion of False Positives.

False Negatives: can be represented as α - TruePositives = α - $\alpha\beta$ = α(1-β)
True Negatives: P(Test - |$\theta = 0$) = P($\theta = 0 \mid Test-) * P(\theta = 0) = \gamma$(1 - α)

c): Once all entries in the table are completed, the solution for C does not necessitate further derivation because the row and column totals are inherently determined by summing across respective rows and columns. While the row totals may appear complex, they are accurate and obtained by computing cell probabilities and summing across rows. Column probabilities, on the other hand, are straightforward to notate, similarly derived from sums within the columns.

(2) ***[20 total points for this part of the problem]*** PPV, NPV, FDR, FOR:

(a) Use Table 2 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*), $P(\theta = 1 \mid y_1 = 1, \mathcal{B})$, and the *negative predictive value* (NPV, with a similar interpretation for negative test results), $P(\theta = 0 \mid y_1 = 0, \mathcal{B})$, of screening tests such as E_1. ***[10 points]***

Solution:
PPV : True Positives / (True Positives + False Positives) = P($\theta = 1 \mid y1 = 1$) = $P(y1 = 1\ \ \&\ \theta = 1)$ * P($\theta = 1$)=P(y1 = 1) = (α β) = ((α β)(1 - α)(1 - γ))

NPV : True Negatives / (True Negatives + False Negatives) = P($\theta = 0 \mid test-) = P(\theta = 0$ and test-) = P(test-) = P(test- |$\theta = 0$) P($\theta = 0$) / P(test-) = γ(1 - α) / (γ(1 - α) + α(1 - β))

(b) 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. ***[10 points]***
Solution: The false discovery rate (FDR) and precision are inversely related, as precision can be calculated as 1 - FDR. Precision indicates the proportion of true positives among all positive test results, while the FDR quantifies the proportion of false positives among all positive test results. FDR is determined by the ratio of false positives to the sum of false positives and true positives, providing insight into the corresponding precision value.

FDR = FP/(FP+TP) Thus it tells us the corresponding value of precision. (1 - α)(1 - γ)=($\alpha\beta$ + (1 - α)(1 - β)).

The false omission rate tells us how often we find a false negative when we predict a negative. This is corresponds to the opposite of our negative predictive value and is equivalent to **1 - NPV**.

(3) ***[30 total points for this part of the problem]*** 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 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.

(a) Use this value for α and the (β, γ) values for E_1 to compute the PPV, NPV, FDR and FOR values defining the blood-screening real-world environment facing the RC in 1985. ***[20 points]***

Solution:

FPR = FP / (FP + TN) = (1 - 0.00286)(1 - 0.95)=((1 - 0.00286)(1 - 0.95) + (1 - 0.00286) * 0.95) = **0.05**

FNR = FN / (FN+TP) = 0.00286(0.01)=(0.00286 * 0.99 + 0.00286(0.01)) = **0.01**

NPV = TN/(TN+FN) = (1 - 0.00286)(0.95)=((1 - 0.00286)*0:95+0.00286*0.01) = **0.9999698093**

PPV = TP/(TP+FP) = 0:00286*0.99=(0.00286*0.99+(1-0.00286)(0.05)) = **0.0537**

(b) Would You say that E_1 was highly successful at keeping HIV out of the RC blood supply in 1985? Explain briefly. ***[10 points]***

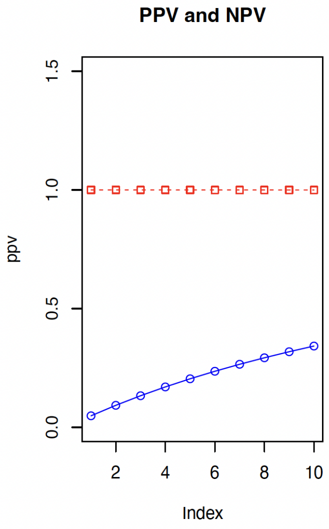
Solution :

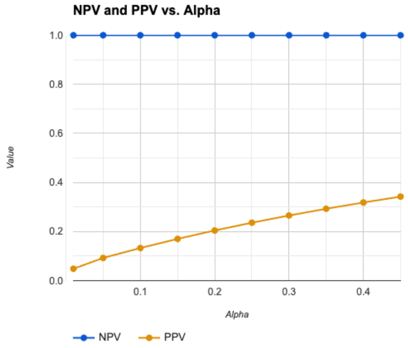
The test demonstrates a notably low false negative rate, indicating its efficacy in minimizing the likelihood of missing true positive cases. However, it concurrently exhibits a limited capacity to detect actual instances, resulting in underutilization of valuable biological samples. The optimization of this trade-off depends on the specific preferences and priorities of the decision-makers involved in the assessment.

(4) ***[20 total points for this part of the problem]*** Sensitivity analysis:

(a) 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, ***with the vertical scale running from 0 to 1***. ***[10 points]***

Solution:

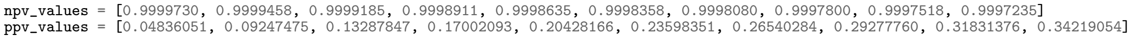




(b) How sensitive were each of these quantities to prevalence in the 1985 RC environment? Explain briefly. ***[10 points]***

Solution:

The negative predictive value (NPV) shown by Red Line of our model exhibits low sensitivity, remaining nearly constant at 1 for all alpha values tested. This suggests the model struggles to reliably identify true negatives, regardless of the chosen significance level. In contrast, the positive predictive value (PPV) displays greater variability, ranging from 0.04 to 0.35 depending on the alpha value. This indicates the model's ability to correctly identify true positives is moderately sensitive to the chosen significance level.


```
npv_values = [0.9999730, 0.9999458, 0.9999185, 0.9998911, 0.9998635, 0.9998358, 0.9998080, 0.9997800, 0.9997518, 0.9997235]
ppv_values = [0.04836051, 0.09247475, 0.13287847, 0.17002093, 0.20428166, 0.23598351, 0.26540284, 0.29277760, 0.31831376, 0.34219054]
```

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 2 with numbers quantifying what would have happened to the $N = 175$ million Americans under this Congress-person'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 2 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 $E(\alpha \,|\, \mathcal{B}) = \alpha^*$.

Table 3: *Partially-filled-out table of expected numbers of people receiving HIV diagnoses under the Congress-person's plan.*

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

(5) ***[40 total points for this part of the problem]*** Real-world implications of the Congress-person's plan:

(a) By looking at the form of all 9 of the entries in Table 2 (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 2 by N and then substituting in $(\alpha, \beta, \gamma) = \left(\frac{1}{350}, 0.99, 0.95\right)$. ***[10 points]***

Solution:

True Positive Cell: Probability of landing in this cell is $\alpha\beta$ ($\alpha =$ unknown, $\beta =$ known)

* **Expected Selling Margin Counts:** We want to estimate these in Table 3 for various α values.

1. **Expected Value:** Due to unknown α, we calculate $E[N\alpha\beta]$ (N and β are constants).
2. We know that : $E[c \cdot X] = c \cdot E[X]$
3. **Conditioning:** $E[\alpha | background] = \alpha^*$ (expected value of α)
4. **Calculation:** Multiply $N\beta\alpha^*$ for each α value to get the corresponding expected selling margin counts in Table 3.

Example:
* N = **175,000,000** - 175 million, $\beta = 0.99$, $\alpha^* = (5/1750)$
Estimated count = $N\beta\alpha^*$

Conclusion: This approach estimates the Expected Selling Margin Counts for various α values, considering the uncertainty in α and leveraging α^* information.

Table 4: *Fully-filled-out table of expected numbers of people receiving HIV diagnoses under the Congress-person's plan.*

		Truth		
		HIV $\oplus$ $(\theta = 1)$	HIV $\ominus$ $(\theta = 0)$	Total
Blood	$\oplus$ $(y_1 = 1)$	495,000	**8,725,000**	**9,220,000**
Test	$\ominus$ $(y_1 = 0)$	**5000**	**165,775,000**	165,780,000
	Total	**500000**	174,500,000	**175000000**

(b) Complete Table 4 by filling in the symbolic cells and margins with the appropriate integers; I've given You a headstart on some of them. $\boxed{\textit{[10 points]}}$

Solution:

Refer table 4 for the reflected numeric values.

We know that:
* N = **175,000,000** - 175 million, $\beta =$ **0.99**, $\alpha^* = (5/1750)$
* TP $= N\beta\alpha = N\beta\alpha^* =$ **495,000**
* $FN = \alpha\,(1-\beta)N \approx$ **5000**
*FP $=$ (1 - α) $(1-\gamma) =$ **8,725,000**
* $TN = (1-\alpha)\,\gamma =$ **165,775,000**
* Total Positive Tests = **9220000**
* True HIV Positive People = **500000**
* True HIV Negative People = **174500000**
* Total Negative Tests = **165780000**
* Total people = **175000000**

(c) Briefly summarize the likely good and bad outcomes of the Congress-person's plan, when viewed as an instance of national health policy. (*Hint:* The first Western Blot test for HIV, which was quite a bit more accurate than E_1, was not developed until 1987.) $\boxed{\textit{[10 points]}}$

Solution:

On one hand, the favorable outcomes entail successfully identifying nearly all individuals with HIV, facilitating prompt treatment. On the other hand, there is the drawback of also identifying approximately 8 million individuals as HIV positive when they do not have the virus, potentially subjecting them to significant inconvenience.

(d) 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.) $\boxed{\textit{[10 points]}}$ **Solution:**

Assessing whether this situation yields a net benefit or a net drawback is challenging, particularly in contexts where individuals bear the cost of their healthcare, such as in the United States. In such circumstances, the prospect of bearing expenses for unnecessary medical interventions could be deemed highly unfavorable.