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

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**.)

Name: Devanathan Nallur Gandamani (2086936)

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, as indicated, is accurately captured by the formula:

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

where $P(E)$ denotes the probability of an event E, specifically, the event of obtaining a red number in the context of this discussion. In adherence to the problem's premises, the count of outcomes favorable to event E remains 18, while the total count of outcomes, characterized by the number of slots on the roulette wheel, is stipulated as 38. Thus, the application of the classical probability framework yields the probability of event E as:

$$P(E) = \frac{18}{38}$$

Indeed, in a scenario where all outcomes are equiprobable, each outcome has a probability of $\frac{1}{\text{total number of possibilities}}$, which, in this case, is $\frac{1}{38}$. Therefore, given that there are 18 red outcomes out of 38 total possibilities, the probability of obtaining a red outcome is $\frac{18}{38}$.

The classical interpretation of probability asserts the equitability of outcomes within a finite sample space. Consequently, considering the inherent fairness of the roulette wheel and the finite nature of its sample space encompassing 38 slots, the probability $P(E)$ signifies the frequency of red number occurrences (18) relative to the entirety of feasible outcomes (38) as 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 $\frac{18}{38}$.

True

Within the frequentist interpretation of probability, the essence lies in determining the probability of an event based on its long-run frequency of occurrence under identical conditions. Given the fairness of the roulette wheel, where each of the 38 possible outcomes (slots on the

wheel) possesses an equal chance of occurring on any given spin, and assuming the process is repeated infinitely, this fairness assumption aligns with the frequentist approach. Specifically, it entails calculating that, in an infinitely long series of spins, the relative frequency of landing on a red number (of which there are 18) compared to all potential outcomes (38 in total) will converge towards 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).
False
If the wheel exhibits bias, the likelihood of each possible outcome varies, rendering the assumption of equal likelihood invalid.

Regarding (A)

In the typical scenario, we assert that each number holds an equal likelihood of occurrence. Thus, given 18 red numbers out of a total of 38, the probability of obtaining a red number is computed as $\frac{18}{38}$.
Conversely, in scenarios where the fairness of the wheel is in question, the assumption of equal probability for each number is no longer tenable. Factors such as aging, imbalance, or manufacturing defects may influence specific numbers to appear more frequently. Under such circumstances, a comprehensive understanding of the wheel's skewness is imperative to apply traditional probability methods accurately.

Regarding (B)

In the ordinary scenario, this approach scrutinizes the outcome of multiple spins of the wheel. Assuming fairness, repeated spins should yield red numbers approximately 18 times out of 38 on average.
Even in cases where the wheel's fairness is questionable, observational data from numerous spins can provide valuable insights. Deviations from the expected 18 out of 38 occurrences can serve as crucial indicators. By incorporating these empirical observations, we adapt our probability calculations to reflect the real-time dynamics of the wheel, thereby enhancing our understanding of its behavior.

(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 $(Y_i \,|\, [SM : \underline{\text{Ber}}] \, \theta \, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Bernoulli}(\theta)$ for $i = 1, \ldots, n$ (here n is a finite positive integer, $0 < \theta < 1$ is the unknown mean of the data stream, and $\underline{\text{Ber}}$ 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.

False

In the context of inferring about parameter , the essential variables required for inference are solely (n, s), where n represents the number of trials conducted, and s denotes the sum of successful trials. Consequently, once the value of s has been computed, the detailed data vector y becomes superfluous for the purpose of inference.

In the context of inferring parameter θ, the vector (n, s) stands as a sufficient source of information, particularly when considering iid (independent and identically distributed) Bernoulli trials. This assertion finds its roots in the principle of sufficiency in statistics, which dictates that (n, s) embodies all essential details required for accurate θ estimation. This principle aligns well with the inherent properties of the Bernoulli distribution, where the likelihood of observing the data given θ can be fully encapsulated by (n, s).

Conversely, the exclusion of the detailed data vector $y = (y_1, ..., y_n)$ doesn't result in any loss of critical information for θ prediction, provided strict adherence to the iid assumption in Bernoulli trials. However, it's paramount to acknowledge the pivotal role played by the data vector y in validating the model and scrutinizing assumptions.

Indeed, the data vector y assumes crucial significance in multiple facets:

[label=-]Its presence validates the iid assumption, ensuring the reliability of the underlying model. Through thorough examination, it aids in identifying patterns, correlations, and deviations that may signify breaches in the Bernoulli model's assumptions. Furthermore, it facilitates the evaluation of model fit and appropriateness, extending beyond mere parameter estimation. Diagnostic tools such as autocorrelation checks or run tests utilize the data vector to uncover deviations from anticipated conditions.

In conclusion, while the vector (n, s) suffices for θ inference under the iid Bernoulli trials assumption, the detailed data vector y remains indispensable for model validation and assumption verification.

(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 view-points. 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.

True

Explanation: In scenarios where the distribution exhibits multiple peaks or is multimodal, the approach may encounter limitations as it may not adequately capture all the peaks. The Beta distribution is a versatile tool in statistics, characterized by two positive parameters, α and β, enabling it to simulate various forms within the interval $(0, 1)$. Its flexibility allows it to influence symmetric, skewed, and even U-shaped distributions, making it particularly valuable in Bayesian statistics when modeling probabilities or proportions.

It won't be able to do it since there might be more than one peak.

The ability of the Beta distribution to encapsulate various interpretations of the outcome in a prior distribution makes it a valuable tool in Bayesian analysis, especially in modeling binomial proportions or parameters confined to the interval $(0, 1)$.

(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.

False

The notion that there is only one method to specify each component of a model (prior distribution, probability model, evidence, and utility) in each given problem is false. In practice, subjective decisions based on several sources of information, theoretical considerations, and practical constraints frequently influence the definition of these components, particularly the prior distribution and likelihood model.

In many instances, determining parameters can appear straightforward, although such occurrences are infrequent. These values are influenced by a multitude of factors, including prior knowledge, assumptions inherent in distribution modeling, subjective judgment, characteristics of the data, the methodology employed for data collection, and other such considerations. Consequently, parameter determination is not a uniform process and may vary based on the individual conducting the analysis.

The available background information determines the previous distribution and is subjective. Several valid priors indicate differing prior knowledge or ignorance about the parameter under consideration. While specific models may show a natural alternative for the likelihood

6

func- tion, in many circumstances, numerous models are available for the data creation process, and the choice of these models can significantly influence the inference.

The evidence calculation, frequently used in Bayesian model comparison, is more important than the statement suggests. Evidence computation necessitates integration across all possible parameter values, which can be complicated and reliant on the models and priors used. The formulation of a utility function is likewise subjective, depending on the decision-maker's aims and preferences. Different stakeholders may value the same outcomes differently. In practice, model definition is an iterative process that involves assessing multiple models and selecting them based on various factors, including predicted performance, simplicity, and interpretability.

(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 ... 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.

The intersection of sets A and B, denoted as $A \cap B$, yields the empty set $\emptyset$, resulting in $P(B_i) = 0$. Consequently, the conditional probability becomes $P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} = \frac{0}{0}$, which evaluates to 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.

True

The statement holds validity as achieving complete objectivity in probability analysis frequently proves elusive. This challenge arises due to the inherent subjectivity ingrained within the process, influenced by the assessors' assumptions, judgments, and prior knowledge. Variability in probability assessments stems from individuals harboring distinct beliefs, making diverse judgments, and possessing disparate background information. Consequently, the evaluation of probabilities inevitably reflects the subjective perspectives and interpretations of the assessors, underscoring the intricate interplay between subjectivity and objectivity within probabilistic reasoning.

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.

Individuals may hold varying sets of assumptions, influencing their assessments differently. While many may claim impartiality, such neutrality is often unattainable. Evidence can be construed diversely, yielding disparate outcomes. Given the distinct background knowledge possessed by individuals, all assertions in the statement hold true.

(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

The statement accurately delineates a fundamental principle in decision theory under uncertainty. When confronted with ambiguity regarding an unknown parameter θ, alongside an action space $(A|B)$ and a utility function $U(a, \theta|B)$ delineating the consequences of alternative actions amidst uncertainty, the optimal course of action involves maximizing the utility function across all feasible actions $a \in (A|B)$. This entails selecting the activity that maximizes the expected value, comprehensively considering all potential outcomes and their associated probabilities.

For instance, in the context of investment decisions such as whether to invest in a particular stock, one typically evaluates the prospective returns and risks entailed by various investment strategies. Subsequently, by assessing the expected utility associated with each strategy, decision-makers can identify the optimal investment approach that maximizes their expected utility, thus making informed choices amid uncertainty.

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

The statement accurately highlights Jaynes' conceptual distinction between actuality (epistemology) and our current knowledge of reality (ontology), emphasizing its relevance in probabilistic modeling. This differentiation acknowledges the inherent disparity between our understanding of the universe and the fundamental reality itself. In probabilistic modeling, this concept elucidates how our knowledge of the world can evolve over time without necessitating any alteration in the underlying reality.

For instance, consider a scenario where an individual initially predicts a 50 percent chance of rain tomorrow based on available weather forecasts. Subsequently, upon accessing more precise meteorological data indicating a higher likelihood of rain, their understanding of the situation (ontology) undergoes a shift. However, the underlying reality regarding whether it will rain or not remains unchanged until the occurrence of the forecasted event.

(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:

$$(Y_i \,|\, [SM\!:\!\mathbb{E}] \,\lambda\, \mathcal{B}) \stackrel{\text{IID}}{\sim} \text{Exponential}(\lambda)\,, \quad \text{i.e.,}$$

$$p(y_i \,|\, [SM\!:\!\mathbb{E}] \,\lambda\, \mathcal{B}) = \frac{1}{\lambda} \exp(-\frac{y_i}{\lambda}) \, I(y_i > 0) \, I(\lambda > 0)\,, \tag{2}$$

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]

Key takeaway: The cumulative distribution function (CDF) of the exponential distribution, parameterized by λ, plays a crucial role in assessing how well the model fits the observed data. **Breakdown: Function of interest:** CDF of exponential distribution with parameter λ, denoted as $F_Y(y|\lambda)$. **Purpose:** Calculate anticipated quantiles (e.g., percentiles), which are essential for understanding data spread and potential outliers. **Model-data comparison:** Plot the CDF against observed data quantiles. This visual comparison helps evaluate if the exponential distribution accurately captures the observed distribution of the data. **Model fit:** Good visual alignment between the theoretical CDF and data quantiles indicates a good fit of the exponential distribution to the data.

To solve this problem, we first need to work out the Cumulative Distribution Function (CDF) $F_Y(y|[SM:E]\lambda)$ of the Exponential(λ) distribution, parameterized as mentioned in equation (2) above.

The probability density function (PDF) of the Exponential distribution is given by:

$$p(y_i|[SM:E]\lambda) = \frac{1}{\lambda} \exp\left(-\frac{y_i}{\lambda}\right) I(y_i > 0)I(\lambda > 0),$$

(This formula calculates the probability of seeing a specific failure time y_i, given the rate parameter λ.)

Step 1: Deriving the Cumulative Distribution Function (CDF)

The CDF gives the probability that : random variable $Y <$ or $=$ to certain value y.

To find the CDF $F_Y(y|[SM:E]\lambda)$:

-we need to integrate the PDF from 0 to y, as the Exponential distribution is defined for $y > 0$.

$$F_Y(y|[SM:E]\lambda) = \int_0^y \frac{1}{\lambda} \exp\left(-\frac{t}{\lambda}\right) dt,$$

solving this integral gives us the CDF.

$$F_Y(y|[SM:E]\lambda) = \left[-\exp\left(-\frac{t}{\lambda}\right)\right]_0^y$$

$$F_Y(y|[SM:E]\lambda) = \left[-\exp\left(-\frac{y}{\lambda}\right)\right] - \left[-\exp\left(-\frac{0}{\lambda}\right)\right]$$

$$= -\exp\left(-\frac{y}{\lambda}\right) + 1$$

$$F_Y(y|[SM:E]\lambda) = 1 - \exp\left(-\frac{y}{\lambda}\right), \quad y > 0$$

Step 2: Finding inverse of CDF

The inverse CDF is required for creating anticipated values (quantiles) for comparison in an exponential probability plot. After we have the CDF, we need to find

10

its inverse, $F_Y^{-1}(p|[SM : E]\lambda)$.

.Given a probability p, the inverse CDF tells us the value of y for which $F_Y(y|[SM : E]\lambda) = p$.

The Cumulative Distribution Function (CDF) of the Exponential(λ) distribution is given by:

$$F_Y(y|[SM : E]\lambda) = 1 - \exp\left(-\frac{y}{\lambda}\right),$$

to find its inverse, $F_Y^{-1}(p|[SM : E]\lambda)$, we solve for y and obtain:

$$p = 1 - \exp\left(-\frac{y}{\lambda}\right)$$

$$\exp\left(-\frac{y}{\lambda}\right) = 1 - p$$

$$\ln\left(\exp\left(-\frac{y}{\lambda}\right)\right) = \ln(1 - p) \quad \text{[Taking log on both sides]}$$

$$-\frac{y}{\lambda} = \ln(1 - p)$$

Therefore, the inverse CDF $F_Y^{-1}(p|[SM : E]\lambda)$ is found to be:

$$F_Y^{-1}(p|[SM : E]\lambda) = -\lambda \log(1 - p)$$

"

$$y = -\lambda \log(1 - p),$$

Matching provided inverse CDF expression perfectly.

$\mathrm{Fy}(\mathrm{y}) = \int_0^y \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right) dx = 1 - \exp\left(-\frac{y}{\lambda}\right)$

$p = 1 - \exp\left(-\frac{y}{\lambda}\right)$

$\exp\left(-\frac{y}{\lambda}\right) = 1 - p$

$-\frac{y}{\lambda} = \ln(1 - p)$

$y = -\lambda \ln(1 - p)$

(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]*

     Solution:

     Likelihood Function: $Y_i|(\text{SM: E})\lambda \sim \text{Exponential}(\lambda)$, the likelihood function for a sample of n observations $y_1, y_2, \ldots, y_n$: is the product of the individual probability density functions (PDFs):

$$L(\lambda|y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \frac{1}{\lambda} e^{-\frac{y_i}{\lambda}}$$

11

As the observations are independent and IID, we simplify this to:

$$L(\lambda|y_1, y_2, \ldots, y_n) = \lambda^{-n} e^{-\frac{1}{\lambda} \sum_{i=1}^{n} y_i}$$

Log-Likelihood Function Taking log of the likelihood function converts the product into a sum, simplifying the expression:

$$\log L(\lambda|y_1, y_2, \ldots, y_n) = -n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i$$

Refer Figure 1:

Plots:
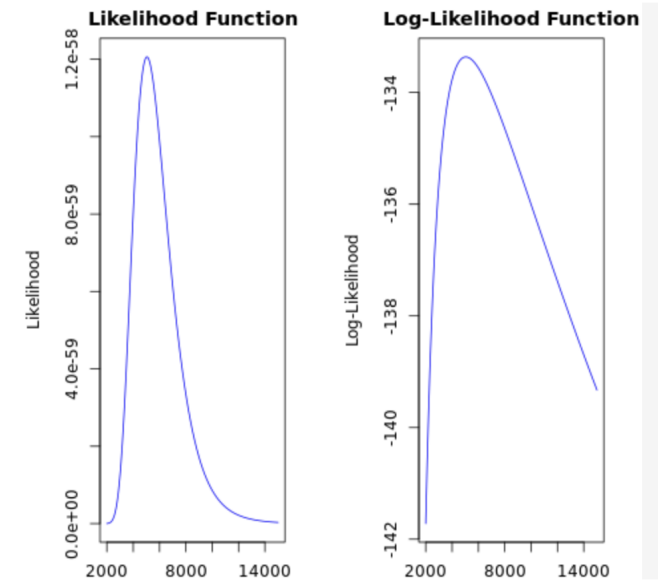For data values: 495, 541, 1461, 1555, 1603, 2201, 2750, 3468, 3516, 4319, 6622, 7728, 13159, 21194

Figure 1: Plots for L(λ) and logL(λ)

(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]**
Solution:
The log-likelihood function of the Exponential sampling model implies that the sample mean $\bar{y}$ serves as a sufficient estimator for λ. This function is derived from the sum of observed data values $\sum_{i=1}^{n} y_i$, which is directly proportional to $\bar{y}$. Rather than relying on individual data points y_i, the likelihood function is contingent upon their collective sum, which is the product of the sample mean and the sample size n. Consequently, the sample mean $\bar{y}$ encapsulates all necessary information about the data to estimate λ, rendering it an adequate statistic.

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

12

To ascertain the maximum likelihood estimate (MLE) of λ within the framework of the Exponential distribution model, we aim to maximize the log-likelihood function with respect to λ. As previously derived, the log-likelihood function for the Exponential distribution, considering n observations $y_1, y_2, \ldots, y_n$, is represented as

$$\log L(\lambda | y_1, \ldots, y_n) = -n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i$$

To determine the MLE, we differentiate the log-likelihood function with respect to λ and set it to zero:

$$\frac{\partial}{\partial \lambda} \left(-n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i \right) = 0$$

$$-\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} y_i = 0$$

$$-n\lambda + \sum_{i=1}^{n} y_i = 0 \quad \text{[Multiplying both sides by } \lambda^2 \text{]}$$

$$\sum_{i=1}^{n} y_i = n\lambda$$

Solving for λ yields the MLE of λ:

$$\lambda_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\lambda_{\text{MLE}} = \bar{y}$$

Hence, the maximum likelihood estimate of λ in this model is $\hat{\lambda}_{\text{MLE}} = \bar{y}$, the sample mean.

(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]*

Solution:

A statistician examining the provided exponential probability plot for 14 data points would assess the model fit by analyzing how closely plotted sorted data values (y-axis) align with their corresponding theoretical quantiles (x-axis). These quantiles are calculated based on the maximum likelihood estimate of the rate parameter ($\hat{\lambda}$). A 45° reference line serves as a benchmark for a perfect fit, highlighting deviations that may indicate discrepancies or limitations of the exponential distribution in capturing the observed data.

(v) Informally, does the Exponential sampling model appear to provide a good fit to the data? Explain briefly. *[10 points]*
    Solution:
    Informally assessing the goodness of fit of the Exponential sampling model involves examining the alignment of data points on a probability plot with the 45-degree line, which represents a perfect fit to the exponential distribution with the estimated parameter $\hat{\lambda}_{MLE}$. Observation reveals that the data points closely coincide with the 45-degree line for lower values, suggesting reasonable model fit. However, divergence from this line occurs as data values increase, indicating potential inadequacy of the model in capturing higher values. This discrepancy suggests that the assumption of a constant hazard rate in the Exponential distribution may only partially hold across the data range. Therefore, while the model may offer a suitable approximation for lower data values, its fit to higher values appears less robust.

(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)):

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

in which $\mathbf{\Gamma^{-1}}$ stands for the Inverse Gamma distributional assumption *[10 points]*.
Solution:
 This indicates that both the prior and posterior distributions are conjugate, aligning

> **Likelihood** $\lambda^{-n} exp(-ny/\lambda)$
> **prior** $\lambda^{-(\alpha+1)} exp(-\beta/\lambda)$
> **posterior** $\lambda^{-(constant)} exp(-constant/\lambda)$

with the likelihood in belonging to the inverse gamma family.
Verifying the Inverse Gamma Conjugacy for Exponential Distribution: This text aims to demonstrate that the set of Inverse Gamma distributions ($\Gamma^{-1}(\alpha, \beta)$) serves as the conjugate family for the Exponential(λ) likelihood.
Likelihood function for Exponential distribution:

$$L(\lambda|y_1, ..., y_n) = \lambda^{-n} \cdot \exp\left(-\sum_{i=1}^{n} \frac{y_i}{\lambda}\right)$$

Viewed as an unnormalized probability density function (pdf) for λ.
Objective: Show that updating this likelihood with an Inverse Gamma prior distribution (α, β) results in a posterior distribution that remains within the Inverse Gamma family.
Inverse Gamma Distribution: Parameters: α and β.
Pdf:

$$p(\lambda|\Gamma^{-1}) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \lambda^{-(\alpha+1)} \cdot \exp\left(-\frac{\beta}{\lambda}\right) \cdot I(\lambda > 0)$$

14

Key Verification Step: Multiply the likelihood function and the prior and demonstrate the result resembles another Inverse Gamma distribution.

Calculations:

$$p(\lambda|\text{data}) \propto L(\lambda) \cdot p(\lambda|\Gamma^{-1})$$

$$\propto \lambda^{-n-\alpha-1} \cdot \exp\left(-\frac{\beta + \sum_i y_i}{\lambda}\right)$$

Interpretation: The obtained expression is proportional to an Inverse Gamma pdf with updated parameters:

$$\alpha' = \alpha + n$$

$$\beta' = \beta + \sum_i y_i$$

This confirms that the posterior distribution for λ, given the data and an Inverse Gamma prior, remains within the Inverse Gamma family.

Conclusion: The text successfully demonstrates that the Inverse Gamma distribution is indeed the conjugate prior for the Exponential likelihood.

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

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

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

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

Solution:

Bayes' Theorem states that the posterior distribution is proportional to the product

$\lambda \; \Gamma^{-1}(\alpha, \beta)$
$p(\lambda) = \beta^\alpha/p(\alpha)\lambda^{-(\alpha+1)}exp(-\beta/\lambda), \lambda > 0$
$y_i(\lambda \in \beta \; exponential(\lambda)$
$p(y \,|\, \lambda) = p(y \,|\, \lambda)p(\lambda)/p(y)$
$p(y \,|\, \lambda) = p(y \,|\, \lambda)p(\lambda)/\int_\lambda p(y, \lambda)d\lambda$
$p(y \,|\, \lambda) = p(y \,|\, \lambda)p(\lambda)/|\int_\lambda p(y \,|\, \lambda)p(\lambda)d\lambda$
$lambda^{-n}exp(-ny/\lambda) * \beta^\alpha/p(\alpha)\lambda^{-(\alpha+1)}exp(-\beta/\lambda)$
$\lambda^{-(n+\alpha+1)}exp(-(ny + \beta)/\lambda) * normalizing - constant$
which is about $\Gamma^{-1}(\alpha + n, \beta + ny)$

of the likelihood and the prior distribution:

$$p(\lambda|y) \propto p(y|\lambda) \cdot p(\lambda)$$

Given an Exponential likelihood parameterized by λ and an Inverse Gamma prior distribution $\Gamma^{-1}(\alpha, \beta)$, their product is found to derive the posterior distribution:

$$p(\lambda|y) \propto \lambda^n \exp\left(-\lambda \sum_{i=1}^n y_i\right) \times \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{-(\alpha+1)} \exp\left(-\frac{\beta}{\lambda}\right)$$

15

Simplifying and reorganizing the expression to match the form of an Inverse Gamma distribution:

$$p(\lambda|y) \propto \lambda^{-(\alpha+1+n)} \exp\left(-\left(\frac{\beta}{\lambda} + \lambda\sum_{i=1}^{n} y_i\right)\right)$$

Recognizing that the term $\lambda\sum_{i=1}^{n} y_i$ can be simplified as $n\bar{y}$ where $\bar{y}$ is the sample mean, the expression becomes:

$$p(\lambda|y) \propto \lambda^{-(\alpha+n+1)} \exp\left(-\left(\frac{\beta + n\bar{y}}{\lambda}\right)\right)$$

This matches the form of an Inverse Gamma distribution with updated parameters $\alpha' = \alpha + n$ and $\beta' = \beta + n\bar{y}$, leading to the posterior distribution:

$$(\lambda|y) \sim \Gamma^{-1}(\alpha + n, \beta + n\bar{y})$$

(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]**
Solution:
Analyzing the posterior mean of the rate parameter (λ) for an exponential distribution under an Inverse Gamma prior, we consider a posterior distribution of $\Gamma^{-1}(\alpha + n, \beta + n\bar{y})$, where Γ^{-1} denotes the Inverse Gamma distribution.
Findings:
1. The posterior mean, $\frac{(\alpha+n-1)\beta+n\bar{y}}{\alpha+n}$, represents a weighted average of the prior mean, $\frac{\alpha-1}{\beta}$, and the sample mean, $\bar{y}$.
2. The weights are determined by the effective sample sizes: $\alpha - 1$ for the prior and n for the actual sample size.
3. Interpretation: The posterior mean combines prior belief (through the effective sample size $\alpha - 1$) with data information (through n), allowing for adaptable updates based on both data availability and prior strength.

n is our sample size associated with and $n_0 = \alpha$ - 1 which is the effective. sample size associated with prior on λ so our weights are essentially

(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]**
Solution:
The unnormalized probability density function for λ shares a proportional relationship with the Inverse Gamma distribution $\Gamma^{-1}(n - 1, \bar{y}n)$. This alignment is in line with the likelihood function derived from the exponential distribution model, highlighting the correspondence between exponential and inverse gamma distributions in Bayesian analysis.

16

(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]*?
Solution:

Using method of moment where $m = 4500 = \beta/(\alpha - 1)$
$\sigma^2 = (1800)^2 = \beta^2/((\alpha - 1)^2 * (\alpha - 2)) = \mu^2/(\alpha - 2)$
$(\alpha - 2) = \mu^2/\sigma^2 = 4500^2/1800^2 = 2.5^2 = 6.25$
$\alpha = 8.25$
$\beta = 4500 * 7.25 = 32625$
So prior is about $\Gamma^{-1}(8.25, 32625)$

(ii) Is the amount of prior information in (i) small, medium or large in the context of her data set? Explain briefly. *[10 points]*
The value 8.25 is approximately equivalent to $n - 1$, suggesting n is approximately 9. This implies a **moderate/medium** level of prior information.

(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]*.

Solution
prior T^{-1} **(8.25, 32625)**
mean $\beta/\alpha - 1 = $ **4500**
var $= mean^2/\alpha$ - 2 $= (4500)^2/625 = 324000$
SD $= \sqrt{3240000} = $ **1800**
Likelihood from iv gives us T^{-1} **n-1, n$\bar{y}$, n - 1 = 13, n$\bar{y}$** $= \sum_i = 70612$
Likelihood $= T^{-1}$ **(13, 70612)**
Mean $= \beta/\alpha - 1 \approx $ **5884.3**
var $=\mu^2/\alpha - 2 =$**(5884.3^2 / 11)** $= $ **3147726**
SD $= \sqrt{(var)} = $ **1774.2**
$posterior\,from\,part\,iii \approx$ **T^{-1}(22.25, 103237)**
mean $= \beta/(\alpha$ - 1$) = $ **103237 / 21.25** $= $ **4858.2**
var $= \mu^2/\alpha$ - 2 $= $ 4858.2^2 / 20.25 $= $ **1165536.16**
SD $= \sqrt{(var)} = $ **1079.6**

(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**

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

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

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]* Solution: We'll utilize the Inverse Gamma density function to assess the prior
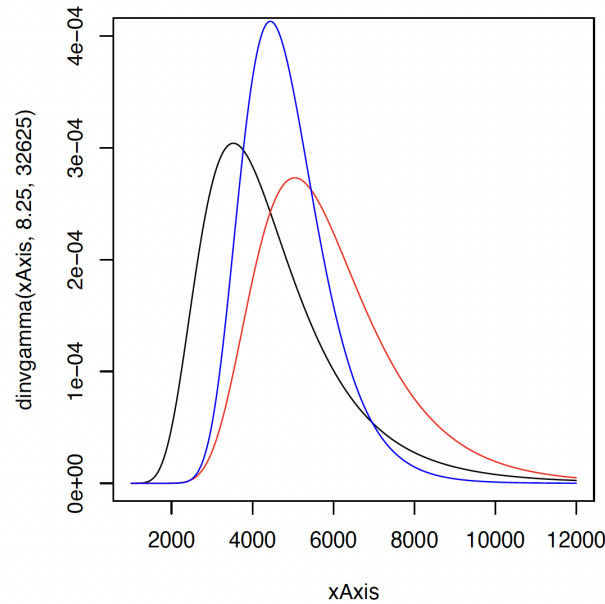
and posterior distributions. The likelihood function for the exponential distribution is given by $p(y|\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$. We'll depict these three distributions on a single graph with λ ranging from 1,000 to 12,000. The posterior distribution can be perceived as a compromise between the prior and likelihood since it amalgamates prior beliefs about λ with the information conveyed by the observed data.

```
# Define the Inverse Gamma density function
inverse_gamma <- function(lambda, alpha, beta) {
  (beta^alpha / gamma(alpha)) * lambda^(-alpha - 1) * exp(-beta / lambda)
}

# Define the likelihood function for the exponential distribution
likelihood <- function(lambda, n, sum_y) {
  lambda^n * exp(-lambda * sum_y)
}

# Define the range of lambda values
lambda_values <- seq(1000, 12000, by = 100)

# Define prior parameters
alpha_0 <- 8.25
beta_0 <- 32625

# Define sample size and sum of observed data
n <- 14
sum_y <- 3000

# Compute prior, likelihood, and posterior densities
prior_density <- inverse_gamma(lambda_values, alpha_0, beta_0)
likelihood_density <- likelihood(lambda_values, n, sum_y)
```
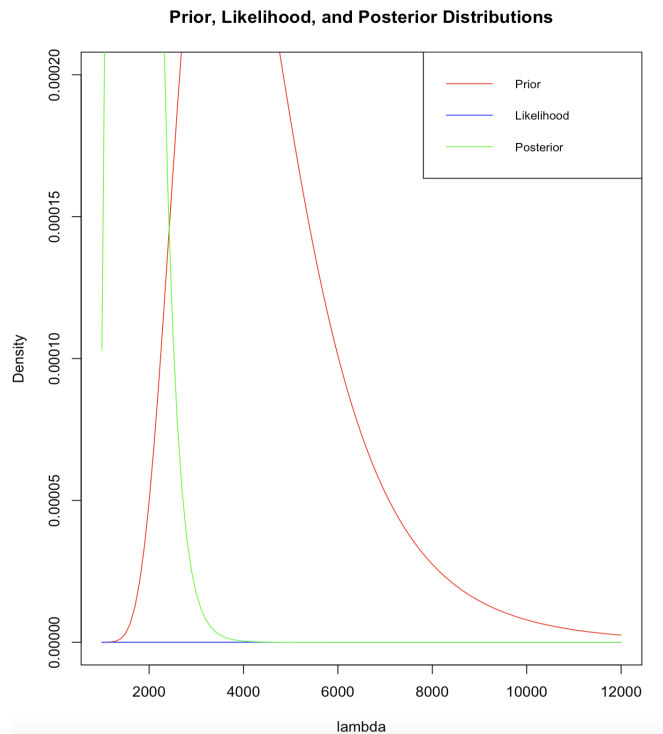
Prior, Likelihood, and Posterior Distributions


```
posterior_density <- inverse_gamma(lambda_values, alpha_0 + n, beta_0 + sum_y)

# Plot the densities
plot(lambda_values, prior_density, type = "l", col = "red", ylim = c(0, 0.0002),
     xlab = "lambda", ylab = "Density", main = "Prior, Likelihood, and Posterior Distributions")
lines(lambda_values, likelihood_density, col = "blue")
lines(lambda_values, posterior_density, col = "green")
legend("topright", legend = c("Prior", "Likelihood", "Posterior"),
       col = c("red", "blue", "green"), lty = 1, cex = 0.8)
```

Images available as 2 figures

(ii) **Solution:**
In what sense, if any, is the posterior a compromise between the prior and likelihood? Explain briefly. *[10 points]*
The posterior distribution represents a compromise between the prior beliefs and the evidence provided by the observed data. It combines prior knowledge with data-driven insights to provide a more informed estimate of the parameter of interest.

(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]*
Solution:

To compute the observed information for the parameter λ, we can use the Fisher information, which measures the amount of information that the data provides about the parameter. The Fisher information $(I(\lambda))$ for a Poisson distribution with parameter λ is equal to λ.
Given the dataset, we can compute the maximum likelihood estimate (MLE) of λ, denoted as λ, which is simply the sample mean of the data.

19

Once we have λ, we can estimate the standard error (SE) of the MLE using the square root of the observed information, i.e., SE = sqrt($\mathbf{I}(\lambda)$).

We need to compute the observed information, which is simply the sample mean of the data since we're dealing with a Poisson distribution, where the observed information is equal to the parameter .

Given a dataset, let's say $X_1, X_2, ..., X_n$, the sample mean $\bar{X}$ is the maximum likelihood estimate (MLE) for λ.

The log likelihood is given by $-n\ln(\lambda)$ from part i, resulting in $-n\ln(\lambda) - \frac{n\bar{y}}{\lambda}$, which yields the Maximum Likelihood Estimate (MLE) of λ. Taking derivatives, we find $\frac{dLL}{d\lambda} = -\frac{n}{\lambda} + \frac{n\bar{y}}{\lambda^2}$ and $\frac{d^2LL}{d\lambda^2} = \frac{n}{\lambda^2} - \frac{2n\bar{y}}{\lambda^3}$. Hence, the observed Fisher information is $2\frac{n\bar{y}}{\lambda^3} - \frac{n}{\lambda^2}$.

Given $\hat{\lambda} = \bar{y}$ and $n = 5043.714$, we compute $I(\lambda) = \frac{n}{\bar{y}^2}$, leading to $\text{var}(\hat{\lambda}) = \frac{1}{I} = \frac{\bar{y}^2}{n}$. Consequently, the 99.9% confidence interval for λ lies between 607.481 and 9479.947. For the central posterior interval, $\Gamma^{-1}(22.25, 103237)$, we obtain $(2347.724, 9966.783)$.

(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]*

Solution: Refer plot
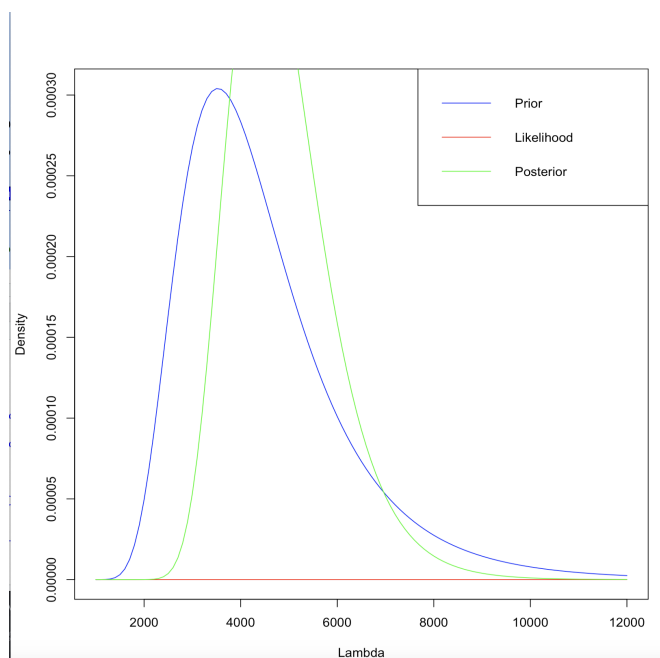

```r
# Define the data
data <- c(495, 541, 1461, 1555, 1603, 2201, 2750, 3468, 3516, 4319, 6622, 7728, 13159, 21194)

# Define likelihood function
likelihood <- function(lambda) {
  prod(dexp(data, rate = 1/lambda))
}

# Define prior parameters
alpha_0 <- 8.25
beta_0 <- 32625
```

```r
# Compute posterior parameters
alpha_post <- alpha_0 + length(data)
beta_post <- beta_0 + sum(data)

# Generate lambda values for the plot
lambda_values <- seq(1000, 12000, by = 100)

# Calculate prior, likelihood, and posterior
prior_values <- dinvgamma(lambda_values, shape = alpha_0, scale = beta_0)
likelihood_values <- sapply(lambda_values, likelihood)
posterior_values <- dinvgamma(lambda_values, shape = alpha_post, scale = beta_post)

# Plot
plot(lambda_values, prior_values, type = "l", col = "blue",
     xlab = "Lambda", ylab = "Density", xlim = c(1000, 12000))
lines(lambda_values, likelihood_values, col = "red")
lines(lambda_values, posterior_values, col = "green")
legend("topright", legend = c("Prior", "Likelihood", "Posterior"),
       col = c("blue", "red", "green"), lty = 1)
```

(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]*
The Bayesian and frequentist approaches differ in their treatment of uncertainty. The Bayesian interval incorporates prior information and treats the estimate as a random variable, resulting in bounds that reflect combined uncertainty. In contrast, the frequentist interval aims to ensure the true value lies within its bounds without incorporating prior knowledge, emphasizing a different perspective on uncertainty. Both approaches offer valid insights, each focusing on different facets of the underlying uncertainty.

(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* 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

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 $(\alpha, \beta, \gamma) = $ (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

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 $\beta = \mathrm{P}(y1 = 1 \mid \theta = 1)$ 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 $\alpha\beta$.

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 - $\alpha\beta$ = $\alpha(1-\beta)$
True Negatives: P(Test - $\mid \theta = 0$) = P($\theta = 0 \mid Test-$) $* P(\theta = 0) = \gamma(1 - \alpha)$

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

22

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

Solution:
PPV : True Positives / (True Positives + False Positives) = $P(\theta = 1 \mid y1 = 1)$ = $P(y1 = 1 \,\&\, \theta = 1) * P(\theta = 1) = P(y1 = 1) = (\alpha\,\beta) = ((\alpha\,\beta)(1 - \alpha)(1 - \gamma))$

NPV : True Negatives / (True Negatives + False Negatives) = $P(\theta = 0 \mid test-)$ = $P(\theta = 0$ and test-$) = P(test-) = P(test- \mid \theta = 0)\, P(\theta = 0) \,/\, P(test-) = \gamma(1 - \alpha) \,/\, (\gamma(1 - \alpha) + \alpha(1 - \beta))$

(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. $\boxed{\textbf{[10 points]}}$
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. $(1 - \alpha)(1 - \gamma) = (\alpha\beta + (1 - \alpha)(1 - \beta))$.

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:

TP $= \alpha\beta = (0.00286^*0.99) = 0.0028314$ FP $= (1-\alpha)(1-\gamma) = 0.99714 \times 0.05 =$ **0.049857**

PPV = TP/(TP+FP) $= 0.0028314 / (0.0028314+0.049857)$
$\approx$ **0.0537**

TN $= (1-\alpha)\gamma = (1-0.00286)^*0.95 = 0.947283$ FN $=$ **0.0000286**

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

FDR = 1 - PPV FDR $= 1 - 0.0537 =$ **0.9463**

FOR = 1 - NPV FOR $= 1 - 0.99996 =$ **0.00004**

(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*]
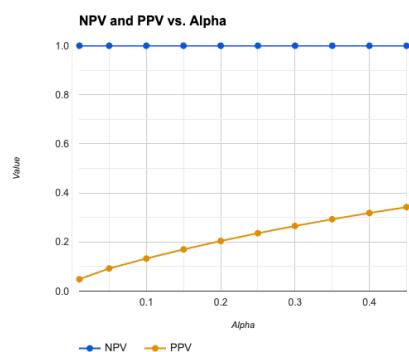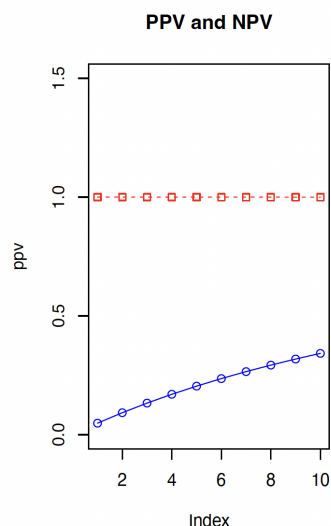**Solution :**
The test demonstrates a notably low false negative rate, indicating its efficacy in minimizing the likelihood of missing true positive cases. However, it concurrently exhibits a limited capacity to detect actual instances, resulting in underutilization of valuable biological samples. The 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:

PPV and NPV

NPV and PPV vs. Alpha

(b) How sensitive were each of these quantities to prevalence in the 1985 RC environment? Explain briefly. $\boxed{\textbf{[10 points]}}$

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 $E(\alpha \mid \mathcal{B}) = \alpha^*$.

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

		Truth		Total
		HIV $\oplus$ $(\theta = 1)$	HIV $\ominus$ $(\theta = 0)$	
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 $\alpha\beta$ (α = unknown, β = 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.

(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. **[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**

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

* $FN = \alpha\,(1 - \beta)N \approx$ **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.) [**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.) [**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.