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 2 ***(625 Total Points)***

(Please see updates in class, by email and `Discord`,
and in `Canvas Announcements` for the **final deadline**.)

Name: Devanathan Nallur Gandamani (2086936)

The (process) ground rules are the same as in Take-Home Test 1: this test is open-book and open-notes, and consists of two problems (true/false and calculation); **each of the 7 true/false questions is worth 10 points, and the calculation problem is worth** $(195 + 360) = 555$ **points, for an overall total of 625 points**.

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

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 (as demonstrated in the LDS meeting ($*$) on Tue 17 Jan 2023); people submitting handwritten solutions can use extra sheets of paper (this was also demonstrated in LDS meeting ($*$)), as long as each solution fragment is clearly marked with the question it's answering.

NB The calculation problems in Section 2 look hard just because they're long, but they're not any harder than usual in this class; because of the extremely compressed nature of this course, I have to do a fair amount of teaching in these problems, just to set up the relevant scientific and statistical questions.

1 True/False

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

(A) Consider the parametric sampling model $(Y_i \,|\, [SM]\, \boldsymbol{\theta}\, \mathcal{B}) \overset{\text{IID}}{\sim} p(y_i \,|\, [SM]\, \boldsymbol{\theta}\, \mathcal{B})$ (for $i = 1, \ldots, n$), where the y_i (the observed values of the Y_i) are real numbers, $\boldsymbol{\theta}$ is a parameter vector of length $1 \leq k \leq n < \infty$ and $\mathcal{B}$ summarizes Your background information; a Bayesian analysis with the same sampling model would add a prior model $[PM]$ layer of the form $(\boldsymbol{\theta} \,|\, [PM]\, \mathcal{B}) \sim p(\boldsymbol{\theta} \,|\, [PM]\, \mathcal{B})$ to the modeling hierarchy. Under mild technical conditions (including regularity as in (D)), the Bernstein-von Mises theorem says that maximum-likelihood (ML) and Bayesian inferential conclusions about $\boldsymbol{\theta}$ will be similar in this setting if (a) n is both large enough and substantially larger than k and (b) $p(\boldsymbol{\theta} \,|\, [PM]\, \mathcal{B})$ is a low-information (LI) prior, but the theorem does not provide guidance on how much bigger n needs to be than k for its conclusion to hold in any specific $[SM]$. ***10 points***

Solution: True

* The Bernstein-von Mises theorem does indeed suggest that ML and Bayesian inference will yield similar results for large n and when p(θ) is diffuse. It is sometimes true because the convergence of the posterior to the normal distribution depends on the satisfaction of several regularity conditions and on the notion of "large n" being sufficient to invoke the central limit theorem-like behavior in the posterior distribution.

* The theorem does not specify how large n needs to be; this is generally determined by the specific context and needs to be assessed on a case-by-case basis. If we were to make the statement universally true, we would have to specify that "for large enough n, under regularity conditions that satisfy the requirements of the Bernstein-von Mises theorem, and assuming a diffuse prior, the Bayesian inference about will be similar to the maximum likelihood approach."

* Additionally, we could clarify that "large enough" is context-dependent and cannot be universally quantified without more information about the specific statistical model and the data.

* As the sample size grows, the influence of the prior distribution on the posterior diminishes, leading to asymptotic properties that dominate as the number of observations approaches infinity. In essence, the predominance of the likelihood function in the posterior composition becomes more pronounced with an increasing sample size, thereby reducing the relative contribution of the prior. Essentially, in the limit, the data overwhelm the prior, such that the latter's impact on posterior conclusions is minimal.

(B) In the basic diagram that illustrates the frequentist inferential paradigm — with the population, sample and repeated-sampling data sets, each containing N, n, and M elements, respectively (see, e.g., page 3.1 of the LDS document camera notes from 25 Jan 2024), and with the sample drawn from the population in an IID manner — when the population parameter of main interest is the mean θ and the estimator is the sample mean $\bar{Y}$, as long as the population SD σ satisfies $(0 < \sigma < \infty)$ You will always get a Gaussian long-run distribution for $\bar{Y}$ (in the repeated-sampling data set) as long as any one of (N, n, M) goes to infinity. ***10 points***

Solution: False

The assertion that increasing the sample size N to infinity, while keeping the number of

parameters M and sample observations n fixed, would alter the long-run distribution is incorrect. The long-run distribution remains unchanged under these conditions. Similarly, the long-run distribution is not impacted by enlarging the number of parameters M while keeping the sample size N and observations n constant. A shift towards convergence is only observed when the number of observations n increases without bound. Therefore, convergence of the long-run distribution is contingent solely on the sample observations n approaching infinity.

(C) The ability to express Your sampling model $[SM]$ as a member of the Exponential Family is helpful, because

- You can then readily identify a set of (minimal) sufficient statistics, and
- a conjugate prior always then exists and can be identified,

in both cases just by looking at the form of the Exponential Family. **10 points**

Solution: True
Indeed, within the Exponential Family, the concept of conjugate priors holds true because the prior-to-posterior updating process results in a posterior distribution that remains within the same family. This facilitates analytical tractability. A conjugate prior is characterized by its ability to maintain the same functional form as the likelihood, enabling a straightforward Bayesian update. Thus, by constructing the conjugate prior to mirror the likelihood's structure, we retain a consistent form during Bayesian inference.

(D) When the sampling model is a regular[1] parametric family $p(\boldsymbol{y} \,|\, [SM]\, \boldsymbol{\theta}\, \mathcal{B})$, where $\boldsymbol{\theta}$ is a vector of length $1 < k \leq n < \infty$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, for n substantially larger than k the repeated-sampling distribution of the (vector) MLE $\hat{\boldsymbol{\theta}}_{MLE}$ is approximately k–variate normal with mean vector $\boldsymbol{\theta}$ and covariance matrix $\hat{I}^{-1}$ (the inverse of the observed information matrix), and the bias of $\hat{\boldsymbol{\theta}}_{MLE}$ as an estimate of $\boldsymbol{\theta}$ in large samples is $O\left(\frac{k}{n^2}\right)$. **10 points**

Solution: False
The bias in the maximum likelihood estimate of $\hat{\theta}$ is not $O\left(\frac{k}{n^2}\right)$, but rather $O\left(\frac{k}{n}\right)$.

(E) It's easier to reason from the part (or the particular, or the sample) to the whole (or the general, or the population), and that's why statistical inference (inductive reasoning) is easier than probability (deductive reasoning). **10 points**

Solution: False
While it is conceptually straightforward to test a specific hypothesis using deductive reasoning, such as verifying a formula, generalizing from a sample to an entire population is inherently more challenging due to the inductive nature of the process. In statistical terms, it's about distinguishing between the ease of hypothesis testing versus the complexities of ensuring sample representativeness. The assertion that a particular sample automatically reflects the broader population is a simplification that disregards the nuanced considerations required for robust statistical generalization. It's akin to the logical fallacy of assuming a stereotype holds true for an entire group based on limited observations, which is a practice fraught with potential for error in both social and statistical contexts.

[1] This means that the ***range*** of possible data values doesn't depend on any components of the parameter vector $\boldsymbol{\theta}$.

(F) When Your sampling model has n observations and a single parameter θ (so that $k = 1$), if the sampling model is regular[1], in large samples the observed information $\hat{I}\left(\hat{\theta}_{MLE}\right)$ is $O(n)$, meaning that

- information in $\hat{\theta}_{MLE}$ about θ increases linearly with n, and
- the repeated-sampling frequentist variance $\hat{V}_F\left(\hat{\theta}_{MLE}\right)$ is $O\left(\frac{1}{n}\right)$.

Solution: True
This is true, there is an inverse relationship between information and variance which is stated.

10 points

(G) One reason that Bayesian inference was not widely used in the early and middle parts of the 20th century was that approximating the (potentially high-dimensional) integrals arising from this approach was difficult in an era when computing was slow (in comparison with contemporary standards) and the Laplace-approximation technique had been forgotten.
10 points

Solution: True
The computational capabilities in the previous century were not adequate for handling the complex integrals that are often encountered in Bayesian inference. The Bayesian methodology necessitates a distinct computational approach when compared to frequentist techniques, which contributed to the lesser focus on Bayesian inference in academic and research publications during that era.

2 Calculation (A)

195 total points From 29–31 Oct 2020, a sample survey was conducted by the highly-regarded polling firm *SurveyUSA*[2] of n = 1,265 adults in the United States who were eligible and likely to vote, to ask about their preferences in the upcoming presidential election. Out of the 1,265 people in the sample, $n_1 = 659$ supported Joe Biden, $n_2 = 554$ supported Donald Trump, and $n_3 = 52$ supported other candidates or expressed no opinion. The polling organization used a sampling method called *stratified random sampling* that's more complicated than the two sampling methods we know about in this class — IID sampling (at random with replacement) and simple random sampling (SRS: at random without replacement) — but here let's pretend that they used SRS from the population $\mathcal{P}$ = {all American people eligible to vote in the U.S. in October 2020 who will actually vote}. There were about 331 million Americans in 2020, of whom about 78% were 18 or older; it was predicted at the time that about 55% of all eligible voters would bother to vote in this election, meaning that $\mathcal{P}$ had about 142 million people in it. The total sample size of n = 1,265 is so small in relation to the population size that we can regard the sampling as effectively IID.

[2] On 2 Nov 2020 the equally high-quality data science website `fivethirtyeight.com` gave the *SurveyUSA* results summarized here a hard-to-get *A* rating, their second highest possible recommendation.

Under these conditions it can be shown, via a generalization of de Finetti's Theorem for binary outcomes, that — since our uncertainty about the responses of the 1,265 people in the survey was exchangeable before the data arrived — the only logically-internally-consistent sampling distribution for the observed data vector $\boldsymbol{n} = (n_1, n_2, n_3)$ is a generalization of the Binomial distribution called the *Multinomial* distribution (You can look back in Your STAT 131 notes, or DeGroot and Schervish (2012), to renew Your acquaintance with the Multinomial).

In a general problem of this type, suppose that a population of interest contains items of $k \geq 2$ types (in the example here: people who support {Biden, Trump, other}, so that in this case $k = 3$) and that the population proportion of items of type j is $0 < \theta_j < 1$. Letting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, note that there's a restriction on the components of $\boldsymbol{\theta}$, namely $\sum_{j=1}^{k} \theta_j = 1$. Now, as in the *SurveyUSA* example, suppose that someone takes an IID sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ of size n from this population and counts how many elements in the sample are of type 1 (call this count n_1), type 2 (n_2), and so on up to type k (n_k); let $\boldsymbol{N} = (N_1, \ldots, N_k)$ be the (vector) random variable that stands for the *process* of getting the data and summarizing it with these counts, and let $\boldsymbol{n} = (n_1, \ldots, n_k)$ be the vector of *observed* counts[3]. In this situation people say that $\boldsymbol{N}$ follows the Multinomial sampling model $[SM\colon \mathbb{Mu}]$ with parameters n and $\boldsymbol{\theta}$, which is defined as follows: $(\boldsymbol{N} \,|\, [SM\colon \mathbb{Mu}]\, n\, \boldsymbol{\theta}\, \mathcal{B}) \sim \text{Multinomial}(n, \boldsymbol{\theta})$ iff

$$P(\boldsymbol{N} = \boldsymbol{n} \,|\, [SM\colon \mathbb{Mu}]\, n\, \boldsymbol{\theta}\, \mathcal{B}) = \left\{ \begin{array}{cc} \frac{n!}{n_1! n_2! \cdots n_k!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_k^{n_k} & \text{if } n_1 + \cdots + n_k = n \\ 0 & \text{otherwise} \end{array} \right\}, \qquad (1)$$

with the further restriction that $0 \leq n_j \leq n$ (for all $j = 1, \ldots, k$). The main scientific and political interest in this problem focuses on $\gamma = (\theta_1 - \theta_2)$, the margin by which Biden was leading Trump on the day of the survey *in the population* $\mathcal{P}$.

The plan in this problem is to work out the likelihood inferential details in parts (a)–(d), to obtain the corresponding Bayesian details in parts (e)–(f), and to summarize Your findings in part (g).

I've written `R` code to help You with the numerical calculations in this part of the test; it's available in the file

`stat-206-tht-2-problem-(2)-(A)-R.txt`

in the `Pages` tab of the course `Canvas` pages.

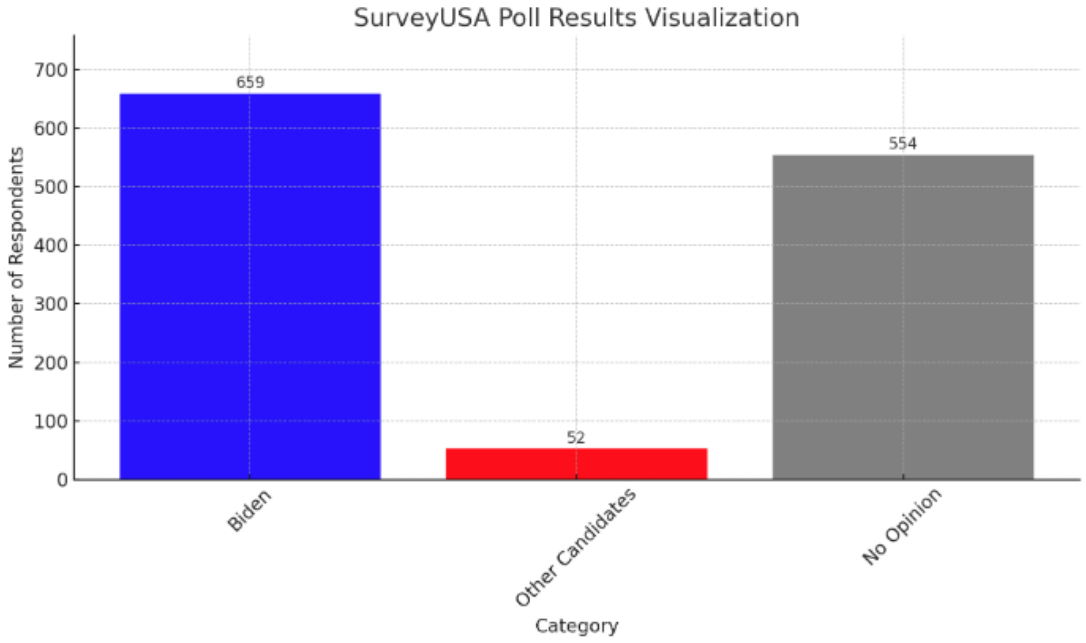
(a) ***15 total points for this part of this problem*** Visualize the raw data set that the *SurveyUSA* people collected, in the form of a data matrix with n rows and 1 column (*Hint:* there are no numbers in this column). Identify all of the following terms (these describe basic data types in data science) that apply to the variable in the single column of Your visualized data set: qualitative, quantitative, categorical, nominal, ordered categorical, dichotomous, discrete, continuous, ratio scale, interval scale. Briefly explain why the numbers $\boldsymbol{n} = (n_1, n_2, n_3) = (659, 554, 52)$ are *not* raw data values but are instead *summaries* of the raw data vector. ***15 points***

Solution:

The raw dataset we're examining consists of individual responses indicating support for either

[3] There is potential notational confusion in this setting that's unavoidable: n is the total sample size here, but $\boldsymbol{n} = (n_1, \ldots, n_k)$ is the observed vector of raw data summaries (note that the latter 'n' is in bold font).

Biden, Trump, or an alternative option, compiled into a list with 1,265 entries corresponding to these choices. This type of data is qualitative in nature and falls into distinct categories, thus it is classified as categorical. Since the sequence in which these responses are presented is irrelevant, the data is also considered nominal. In essence, the dataset is characterized by qualitative, categorical, and nominal attributes. The figures 659, 554, and 52 are not the raw data themselves but are summary counts reflecting the number of occurrences within each of the three respective categories: 659 entries for Biden, 554 for alternative options, and 52 for Trump.



(b) ***10 total points for this part of this problem*** Show that the Multinomial is indeed a direct generalization of the Binomial, if we're careful in the notational conventions we adopt. Here's what I mean: as You know, the Binomial sampling model $[SM\colon \mathbb{B}]$ arises when somebody makes n IID success–failure (Bernoulli) trials, each with success probability θ, and records the number X of successes; this yields the sampling distribution

$$(X \,|\, [SM\colon \mathbb{B}]\, n\, \theta\, \mathcal{B}) \sim \text{Binomial}(n, \theta) \quad \text{iff}$$

$$P(X = x \,|\, [SM\colon \mathbb{B}]\, n\, \theta\, \mathcal{B}) \quad = \quad \left\{ \begin{array}{cc} \binom{n}{x} \theta^x (1-\theta)^{n-x} & \text{for } x = 0, \ldots, n \\ 0 & \text{otherwise} \end{array} \right\} \quad (2)$$

Briefly and carefully explain why the correspondence between equation (2) and {a version of equation (1) with $k = 2$} is as in Table 1. ***10 points***

Two comments are worth making here:

- The Multinomial PMF has something interesting hidden inside it: suppose that we wanted to combine two of the three categories {Biden, Trump, Other}, e.g., to create {Biden, Not-Biden}; the result would be a new Multinomial PMF in which everything is logically

Table 1: *The Binomial as a special case of the Multinomial: notational correspondence.*

Binomial	Multinomial ($k = 2$)
n	n
x	n_1
$(n - x)$	n_2
θ	θ_1
$(1 - \theta)$	θ_2

internally consistent with the original Multinomial (e.g., the new n for {Not-Biden} would be the sum of the old n values for {Trump} and {Other}, and the new θ for {Not-Biden} would be the sum of the old θ values for {Trump} and {Other}). Natural first reaction to this: that's cool; natural second reaction: if that ***didn't*** work, something would be wrong.

- Following on from (a) above, let Y_i record the voting preference for sampled person i, coded as one of the character strings $\boldsymbol{C} \triangleq$ {'Biden', 'Trump', 'Other'}, in that order; then the components Y_i of the raw data vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ follow what's called a ***Categorical (PMF)*** $[SM : \mathbb{C}\text{a}]$, which differs from the distributions of all of the random variables we studied in STAT 131 in that *the values of the Y_i are not real numbers*:

$$\left\{ \begin{array}{c} (Y_i \,|\, [SM : \mathbb{C}\text{a}]\, \boldsymbol{C}\, \boldsymbol{\theta}\, \mathcal{B}) \overset{\text{IID}}{\sim} \\ \underline{\text{Ca}}\text{tegorical}(\boldsymbol{C}, \boldsymbol{\theta}) \end{array} \right\} \longleftrightarrow p(y_i \,|\, [SM : \mathbb{C}\text{a}]\, \boldsymbol{C}\, \boldsymbol{\theta}\, \mathcal{B}) = \left\{ \begin{array}{lcc} \theta_1 & \text{if} & y_i = \text{'Biden'} \\ \theta_2 & & y_i = \text{'Trump'} \\ \theta_3 & & y_i = \text{'Other'} \\ 0 & & \text{otherwise} \end{array} \right\}, \tag{3}$$

with $0 < \theta_j < 1$ and $\sum_{j=1}^{3} \theta_j = 1$. It's easy to show that the vector $\boldsymbol{N} = (N_1, N_2, N_3)$ forms a set of (minimal) sufficient statistics for the vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in this sampling model, and one of the consequences of the likelihood story is that, given this sufficient-statistic result,

> We can build our likelihood function for $\boldsymbol{\theta}$ either directly from equation (3) or from the Multinomial $[SM]$ for $\boldsymbol{N}$, and we'll get the same results either way: this is called ***reduction by sufficiency (from $\boldsymbol{Y}$ to $\boldsymbol{N}$)***.

In what follows we'll work directly with $\boldsymbol{N}$, using sufficiency to park $\boldsymbol{Y}$ on the sidelines.

Solution:

* The Binomial distribution is essentially a constrained Multinomial distribution with two outcomes, which is to say that the Binomial can be viewed as a binarized instantiation of the Multinomial. This relationship becomes evident when considering the structure of IID Bernoulli trials, which are characterized by binary success-failure outcomes. In this framework, the Binomial distribution enumerates the probability of accruing a certain number of successes across a series of trials, where success occurs with probability . Within the purview of the Multinomial distribution, when k equals 2, this translates to a scenario where the outcome space is bifurcated into successes and failures, each with associated probabilities that mirror those in the Binomial case. Thus, the parameters and probabilities of the Binomial are directly mapped onto a two-outcome Multinomial context, reflecting its foundational role as a special case within the broader Multinomial paradigm. n = n as it represents the

sample size in both multinomial and binomial representations
x = n_1 we can relate these as in the binomial we use θ^x while in the binomial representation we have θ^{n_1} we can then infer that both of these are related as what we're searching for
(n - x) = n_2 We can see that these items are equivalent, we can see that for our second success probability, when equating the two formulas and accounting for what we've already done that these two terms are equivalent
$\theta = \theta_1$ These represent both our initial success probabilities in the binomial and multinomial distribution.
(1 - θ) = θ_2 as the last remaining term, these terms represent the same thing for our binomial and multinomial formulas.

** The Multinomial distribution generalizes the Binomial distribution to the case where there are more than two outcomes. Specifically, when $k = 2$, the Multinomial distribution reduces to the Binomial distribution.
*** Consider a Multinomial distribution with two outcomes, such that $\theta_1 + \theta_2 = 1$ and $n_1 + n_2 = n$, where n is the total number of trials, n_1 is the number of successes, and n_2 is the number of failures. If we let $\theta_1 = \theta$ and $\theta_2 = 1 - \theta$, then the probability mass function for the Multinomial distribution simplifies to:

$$\begin{aligned} P(X = x \mid \text{SM: } n\theta) &= \binom{n}{x} \theta_1^{n_1} \theta_2^{n_2} \\ &= \binom{n}{x} \theta^{n_1} (1-\theta)^{n_2} \\ &= \binom{n}{x} \theta^{x} (1-\theta)^{n-x} \end{aligned}$$

which is exactly the probability mass function of a Binomial distribution with parameters n and θ.

(c) ***40 total points for this part of this problem*** Returning now to the general Multinomial setting:

(i) Briefly explain why the likelihood function for $\boldsymbol{\theta}$ given the observed vector $\boldsymbol{n}$ of data summaries and $\mathcal{B}$ is

$$\ell_C(\boldsymbol{\theta} \mid [SM\text{: } \mathbb{M}\mathbb{u}]\, \boldsymbol{n}\, \mathcal{B}) = c_+ \prod_{j=1}^{k} \theta_j^{n_j} \tag{4}$$

(in which c_+ is, as usual, an arbitrary positive constant), leading to the log-likelihood function

$$\ell\ell_C(\boldsymbol{\theta} \mid [SM\text{: } \mathbb{M}\mathbb{u}]\, \boldsymbol{n}\, \mathcal{B}) = c + \sum_{j=1}^{k} n_j \log \theta_j \,, \tag{5}$$

where c is an arbitrary real constant. ***[10 points]***

Solution:
The likelihood function mirrors our joint sampling distribution for $N = (N_1, ..., N_k)$, represented as $p(N = n|\theta B) = c\prod_{j=1}^{k} \theta_j^{n_j}$, under the condition that the sum of θ_j equals one. Consequently, the log likelihood function, derived by taking the logarithm of each term and aggregating them, instead of their products, is expressed as $(\theta|nB) = \sum_{j=1}^{k} n_j \log \theta_j$. This approach is essentially a log-transformation of the joint probability distribution.

In finding the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, if You simply try, as usual, to set all of the first partial derivatives of $\ell\ell_C(\boldsymbol{\theta} \,|\, [SM\colon \mathbb{M}\mathfrak{u}]\, \boldsymbol{n}\, \mathcal{B})$ with respect to the θ_j equal to 0, You'll get a system of equations that has no solution (try it). This is because in so doing we forgot that we need to do a *constrained optimization*, in which the constraint is $\sum_{j=1}^{k} \theta_j = 1$ (this explains the subscript C in equations (4) and (5): it stands for *Constrained*). There are thus two ways forward to compute the MLE (You're requested to perform both computations):

(ii) Solve the constrained optimization problem directly with *Lagrange multipliers* (The TAs and I will show you how to do this in office hours if You forget or don't know, because `Wolfram Alpha` seems to be useless here) ***[10 points]***,

Solution:
Our log likelihood = $LL = \lambda(\sum_{j=1}^{k}(\theta_j - 1)) = 0$
If we take partial derivatives of both sides, $d/d\theta_j$, we find a relation that every θ, $\theta_1, \theta_2, ..., \theta_k$ we get an equivalency such that $n_k/\theta_k = \lambda$
We can then solve for θ_k showing that for any $\theta_k = n_1/\lambda$
Because of the constraint where all our thetas sum to one, we get $1 = n/\lambda$ or $\lambda = n$
Which means that our MLE's $\theta = (n_1/n, n_2/n, ..., n_k/n)$ so long as $\sum_j n_j = n$

(iii) Build the constraint directly into the likelihood function: since $\sum_{j=1}^{k} \theta_j = 1$, we can arbitrarily pick one of the θ_j, say θ_k, and write it as $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$; then, defining $\boldsymbol{\theta}_{[-k]} \triangleq (\theta_1, \ldots, \theta_{k-1})$ (i.e., $\boldsymbol{\theta}_{[-k]}$ is $\boldsymbol{\theta}$ with θ_k omitted), we can set

$$\ell_U(\boldsymbol{\theta}_{[-k]} \,|\, [SM\colon \mathbb{M}\mathfrak{u}]\, \boldsymbol{n}\, \mathcal{B}) \triangleq c_+ \left(\prod_{j=1}^{k-1} \theta_j^{n_j}\right) \left(1 - \sum_{j=1}^{k-1} \theta_j\right)^{n_k} \tag{6}$$

(here the subscript U stands for *Unconstrained*). At this point it becomes helpful in the algebra that follows to substitute $\left(n - \sum_{j=1}^{k-1} n_j\right)$ for n_k and take logs to obtain

$$\begin{aligned} \ell\ell_U(\boldsymbol{\theta}_{[-k]} \,|\, [SM\colon \mathbb{M}\mathfrak{u}]\, \boldsymbol{n}\, \mathcal{B}) \;\; &= \;\; c + \sum_{j=1}^{k-1} n_j \, \log \theta_j \\ &\quad + \left(n - \sum_{j=1}^{k-1} n_j\right) \log \left(1 - \sum_{j=1}^{k-1} \theta_j\right) . \end{aligned} \tag{7}$$

For $j = 1, \ldots, (k-1)$, show that

$$\frac{\partial}{\partial \theta_j} \ell\ell_U(\boldsymbol{\theta}_{[-k]} \,|\, [SM\colon \mathbb{M}\mathfrak{u}]\, \boldsymbol{n}\, \mathcal{B}) = \frac{n_j}{\theta_j} - \frac{n - \sum_{j=1}^{k-1} n_j}{1 - \sum_{i=1}^{k-1} \theta_i} \tag{8}$$

[10 points]

Solution:

p1

Given the probabilities $(\theta_1, \ldots, \theta_{k-1}|nB)$, the likelihood is proportional to:

$$c \left(\prod_{j=1}^{k-1} \theta_j^{n_j} \right) \left(1 - \sum_{j=1}^{k-1} \theta_j \right)^{n_k}$$

The log likelihood function is:

$$\sum_{j=1}^{k-1} n_j \log \theta_j + n_k \log \left(1 - \sum_{j=1}^{k-1} \theta_j \right)$$

Deriving the log likelihood with respect to θ_j yields:

$$\frac{dLL}{d\theta_j} = \frac{n_j}{\theta_j} + n_k \left(\frac{-1}{1 - \sum_{i=1}^{k-1} \theta_i} \right)$$

for $j = 1, \ldots, k-1$.

p2

The first-order condition for optimization, $\frac{n_j}{\theta_j} - n_k \left(\frac{1}{1 - \sum_{i=1}^{k-1} \theta_i} \right) = 0$, leads to:

$$\hat{\theta_j} = \frac{n_j (1 - \sum_{i=1}^{k-1} \theta_i)}{n_k}$$

By utilizing the constraint that the sum of all θ_j equals one, we find:

$$\theta_k (n(n - n_k)/n_k + 1) = 1$$

Solving for θ_k gives us $\hat{\theta_k} = n_k/n$. Substituting j for k in the equations, we find $\hat{\theta_j} = n_j/n$ for $j = 1, \ldots, k-1$.

p3

When $k = 3$, we have the equations:

$$\frac{n1}{\theta_1} - \frac{n3(1 - \theta_1 - \theta_2)}{0} = 0$$

$$\frac{n2}{\theta_2} - \frac{n3(1 - \theta_1 - \theta_2)}{0} = 0$$

Solving for θ gives us $\theta_1 = n_1/n_3$ and $\theta_2 = n_2/n_3$.

Expanding the equations yields:

$$\theta_1(1 + \frac{n1}{n3}) + \theta_2(\frac{n1}{n3}) = \frac{n1}{n3}$$

$$\theta_2(1 + \frac{n2}{n3}) + \theta_1(\frac{n2}{n3}) = \frac{n2}{n3}$$

Let $a = \frac{n1}{n3}$ and $b = \frac{n2}{n3}$. Then $\theta_1 = a/(a + b + 1)$ and $\theta_2 = b/(a + b + 1)$.

1. **Derivation of Log-likelihood Function (p1):**
   - It starts by expressing the likelihood function for a multinomial distribution, which is a product of the probabilities raised to the number of occurrences. The likelihood is then converted into a log-likelihood function by taking the logarithm of each term and summing them. This is useful because it simplifies the multiplication of probabilities into a sum, which is easier to work with mathematically.
2. **Partial Derivation and Constraint Application (p2):**
   - To find the MLEs, partial derivatives of the log-likelihood function with respect to each θ_j are taken. This is done to find the points where the log-likelihood function is maximized.
   - There is a constraint that all the probabilities θ_j must sum to one. This constraint is incorporated into the optimization problem using a method such as Lagrange multipliers.
3. **Solution for θ_j (p3):**
   - A system of equations is created based on the partial derivatives set to zero. The equations are then solved for θ_1 and θ_2, showing that each parameter θ_j is equal to the ratio of the number of occurrences of outcome j to the total number of occurrences n.
4. **Specific Case of $k = 3$:**
   - For the case where there are three outcomes (k=3), a specific system of equations is solved to express θ_1 and θ_2 in terms of n_1, n_2, and n_3. This results in formulas that give the MLEs of the parameters in terms of the sample proportions.

The MLE for $(\theta_1, \ldots, \theta_{k-1})$ may now be found by setting

$$\frac{\partial}{\partial\, \theta_j}\, \ell\ell_U(\theta_1, \ldots, \theta_{k-1} \,|\, [SM\colon \mathbb{M}\mathbb{u}]\, \boldsymbol{n}\, \mathcal{B}) = 0 \quad \text{for } j = 1, \ldots, (k-1) \tag{9}$$

and solving the resulting system of $(k-1)$ equations in $(k-1)$ unknowns, but that gets quite messy; let's just do it for $k = 3$, which is all we need in the *SurveyUSA* context anyway.

(iv) Solve the two equations

$$\left\{ \frac{n_1}{\theta_1} - \frac{n - n_1 - n_2}{1 - \theta_1 - \theta_2} = 0, \quad \frac{n_2}{\theta_2} - \frac{n - n_1 - n_2}{1 - \theta_1 - \theta_2} = 0 \right\} \tag{10}$$

for (θ_1, θ_2) and then use the constraints $\sum_{j=1}^{3} \theta_j = 1$ and $\sum_{j=1}^{3} n_j = n$ to get the MLE for θ_3, thereby demonstrating the (entirely obvious, after the fact) result that

$$\hat{\boldsymbol{\theta}}_{MLE} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}\right) . \tag{11}$$

[10 points] (The result for general k, of course, is[4] that $\hat{\boldsymbol{\theta}}_{MLE} = \frac{1}{n}\boldsymbol{N}$. With $\gamma = (\theta_1 - \theta_2)$ defined as above, note that, by functional invariance of the MLE, $\hat{\gamma}_{MLE} = (\hat{\theta}_1 - \hat{\theta}_2)$.)

Solution:

To solve the system of equations for θ_1 and θ_2 given in the image, we use the following steps:

Given equations:

$$\begin{cases} \frac{n_1}{\theta_1} = \frac{n - n_1 - n_2}{1 - \theta_1 - \theta_2} \\ \frac{n_2}{\theta_2} = \frac{n - n_1 - n_2}{1 - \theta_1 - \theta_2} \end{cases}$$

Adding the constraints that:

$$\sum_{j=1}^{3} \theta_j = 1 \text{ and } \sum_{j=1}^{3} n_j = n$$

We know that:

$$\theta_3 = 1 - \theta_1 - \theta_2$$

From the constraints, we also have that $n_3 = n - n_1 - n_2$.

By solving the system, we aim to find expressions for θ_1 and θ_2 in terms of n_1, n_2, n_3, and n. We can then express θ_3 as n_3/n because it must account for the remaining proportion of the total n.

Since the system of equations are derived from the condition that the ratio of observed counts over their respective probabilities must be equal, we can equate the two ratios:

$$\frac{n_1}{\theta_1} = \frac{n_2}{\theta_2} = \frac{n - n_1 - n_2}{1 - \theta_1 - \theta_2} = \frac{n_3}{\theta_3}$$

Thus, the maximum likelihood estimates (MLE) for $\theta_1, \theta_2, \theta_3$ can be obtained as:

$$\hat{\theta}_1 = \frac{n_1}{n}, \quad \hat{\theta}_2 = \frac{n_2}{n}, \quad \hat{\theta}_3 = \frac{n_3}{n}$$

[4] To conform to the notational conventions in this course, I should write $\hat{\boldsymbol{\Theta}}_{MLE} = (\hat{\Theta}_1, \ldots, \hat{\Theta}_k) = \frac{1}{n}\boldsymbol{N}$ instead of $\hat{\boldsymbol{\theta}}_{MLE} = (\hat{\theta}_1, \ldots \hat{\theta}_k)$, using capital letters to denote random variables and lower-case letters to stand for their possible values, but the result in this problem gets quite ugly if I do so; I will also sometimes drop the subscript *MLE* and just go, e.g., with $\hat{\theta}_1$; please note and excuse these departures from otherwise common practice in this class.

The result demonstrates the intuitive solution that the MLEs for the parameters of a multinomial distribution are simply the proportions of the counts in each category.
Finally, the general result for any k is that $\hat{\theta}_{MLE} = \frac{1}{n}N$, where N is the vector of counts for each category.
For $\gamma = (\theta_1 - \theta_2)$, by the functional invariance of the MLE, we have $\hat{\gamma}_{MLE} = \hat{\theta}_1 - \hat{\theta}_2$.

(d) ***40 total points for this part of this problem*** Run **Code Block 1** in the `R` code file mentioned at the beginning of this problem; study the results, and use them in Your answers to the questions in (d)(i) and (d)(ii) below.

It can be shown (You're not asked to show this) that in repeated sampling (with $k = 3$) the estimated covariance matrix of the MLE vector $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right)$ is

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n} & -\frac{\hat{\theta}_1\,\hat{\theta}_2}{n} & -\frac{\hat{\theta}_1\,\hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1\,\hat{\theta}_2}{n} & \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n} & -\frac{\hat{\theta}_2\,\hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1\,\hat{\theta}_3}{n} & -\frac{\hat{\theta}_2\,\hat{\theta}_3}{n} & \frac{\hat{\theta}_3(1-\hat{\theta}_3)}{n} \end{pmatrix} . \tag{12}$$

(i) Use $\hat{\boldsymbol{\Sigma}}$ to compute approximate large-sample standard errors for the MLEs of the θ_i and of γ; for $\widehat{SE}(\hat{\gamma})$ You can either (You're not requested to do both)

* work out $\widehat{SE}(\hat{\gamma})$ directly, by thinking about the repeated-sampling variance of the difference of two (correlated) random quantities, or
* use the fact (from STAT 131) that if $\hat{\boldsymbol{\theta}}$ is a $(k \times 1)$ random vector with covariance matrix $\hat{\boldsymbol{\Sigma}}$ and $\gamma = \boldsymbol{a}^T\boldsymbol{\theta}$ for some $(k \times 1)$ vector $\boldsymbol{a}$ of constants, then in repeated sampling

$$\hat{V}(\hat{\gamma}) = \hat{V}\left(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}\right) = \boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}\,\boldsymbol{a}\,. \tag{13}$$

[10 points]

(9A)

(ds(i)

Formula to calculate covariance matrix MLE

$$\Sigma = \begin{bmatrix} -1.972815 \times 10^{-4} & -1.803531 \times 10^{-4} & -1.692849 \times 10^{-5} \\ -1.803531 \times 10^{-4} & -1.948883 \times 10^{-4} & -1.423120 \times 10^{-5} \\ -1.692849 \times 10^{-5} & -1.423120 \times 10^{-5} & 3.115965 \times 10^{-5} \end{bmatrix}$$

$\because \gamma = \hat{\theta}_1 - \hat{\theta}_2$

$\hat{V}(\hat{\gamma}) = \hat{V}(\hat{\theta}_1) + \hat{V}(\hat{\theta}_2) - 2 * \text{Covariance}(\hat{\theta}_1, \hat{\theta}_2)$

$\hat{V}(\hat{\gamma}) = 0.000619728 + 0.000194584 - 2 * (0.000180531)$

$= 0.00074526$

$\widehat{SE}(\hat{\gamma}) = \sqrt{\hat{V}(\hat{\gamma})} = 0.027432827 05$

As noted above, the principal scientific and political interest here is the amount γ by which Mr. Biden was leading Trump at the time of the *SurveyUSA* poll; a Devil's Advocate (DA) would say (I) that $\gamma = 0$ and (II) that the only reason the survey got a positive estimate of γ was unlucky random sampling. To judge the plausibility of the DA's claim we need a modification of Mr. Neyman's confidence-interval machinery called a ***(one-sided) lower confidence bound (LCB)*** for γ. It can be shown (You're not asked to show this) that

$$\hat{\gamma}_{MLE} - \Phi^{-1}(1-\alpha) \cdot \widehat{SE}\left(\hat{\gamma}_{MLE}\right) \tag{14}$$

is an approximate $100\,(1-\alpha)\%$ LCB for γ; in other words, we're $100\,(1-\alpha)\%$ confident that γ is *at least* equal to the value in equation (14).

(ii) Is Mr. Biden's lead practically significant? Statistically significant? Let's see.

(∗) Was Mr. Biden ahead of Trump at the point when the survey was conducted by an amount that was large in *practical* terms? Explain briefly. ***[10 points]***

Solution:
The lower bound is negative, which indicates statistically Biden's lead is significant. However, in practical situations, especially considering the use of the electoral vote in the USA, the lead may not directly translate into who wins the election.

(**) Use the relevant output from **Code Block 2** to announce an approximate (large-sample) 99.9% LCB for γ Was Biden's lead at that point *statistically* significant at the 99.9% level? Explain briefly. ***[10 points]***

Solution:

The output of the code to calculate the 99.9% LCB for V is -0.0017; this means we are 99.9% sure that Biden's lead at that point was -0.17%.
This means that if Biden were to lose, the margin can't be more than 0.17% of Trump's vote, and we can be 99.9% sure of this.

(***) Repeat (**) with a 99.8% confidence level. What does this say about the concept of **statsig**? Explain briefly. ***[10 points]***

Solution:
The output of the code to calculate the 99.8% Lower Confidence Bound (LCB) for V is 0.00404. This means we are 99.8% sure that Biden's lead at that point was 0.404% at least. That is, we can be 99.8% sure about Biden's win. We can conclude that Biden's lead is significant statistically, as the lower bound is positive. This opposes the null hypothesis that both have an equal chance at winning.

(e) ***20 total points for this part of this problem*** Looking back at equation (4), if a conjugate prior exists for the Multinomial likelihood it would have to be of the form

> c_+ times θ_1 to a power times θ_2 to a (possibly different) power times ... times θ_k to a (possibly different) power.

There is such a distribution — it's called the $\mathbb{D}$**irichlet**$(\boldsymbol{\alpha})$ distribution (You can learn more about it in *Appendix A* of the Gelman et al. book)), with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ chosen so that all of the α_j are positive and finite:

$$p(\boldsymbol{\theta} \,|\, [PM\colon \mathbb{D}]\, \boldsymbol{\alpha}\, \mathcal{B}) = c_+ \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}\,; \tag{15}$$

note that, as usual with conjugate priors, the Dirichlet $[PM]$ assumption is not part of $\mathcal{B}$. The Dirichlet distribution is a multivariate generalization of the Beta(α, β) distribution; as we saw in Table 1 , to see the correspondence you just have to replace $[\theta, (1-\theta)]$ with (θ_1, θ_2) and (α, β) with (α_1, α_2).

(i) Briefly explain why this means that the conjugate updating rule is

$$\left\{ \begin{array}{ccc} (\boldsymbol{\theta} \,|\, [PM\colon \mathbb{D}]\, \boldsymbol{\alpha}\, \mathcal{B}) & \sim & \text{Dirichlet}(\boldsymbol{\alpha}) \\ (\boldsymbol{N} \,|\, [SM\colon \mathbb{Mu}]\, \boldsymbol{n}\, \boldsymbol{\theta}\, \mathcal{B}) & \sim & \text{Multinomial}(\boldsymbol{n}, \boldsymbol{\theta}) \end{array} \right\} \longrightarrow$$
$$(\boldsymbol{\theta} \,|\, [PM\colon \mathbb{D}]\, \boldsymbol{\alpha}\, [SM\colon \mathbb{Mu}]\, \boldsymbol{n}\, \mathcal{B}) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{n})\,, \tag{16}$$

in which $(\boldsymbol{\alpha} + \boldsymbol{n})$ is a vector sum. ***[10 points]***

Solution:

$$\Theta|(\text{PM} : D) \propto B = c_+ \prod_{j=1}^{k} \Theta_j^{\alpha_j - 1}$$

Sampling model is multinomial.

$$(N|\text{SM} : M \cup \cap B\theta) \propto \prod_{j=1}^{k} \Theta_j^{n_j}$$

The posterior $\propto$ prior $\times$ likelihood

$$(\Theta|\text{PM} : D) \propto B = c_+ \prod_{j=1}^{k} \Theta_j^{\alpha_j - 1} \prod_{j=1}^{k} \Theta_j^{n_j - 1}$$

or Dirichlet$(\alpha + N_j)$ or Dirichlet$(\alpha + N)$.

(ii) Given that $\boldsymbol{N} = (n_1, \ldots, n_k)$ and that the n_j represent sample sizes (numbers of observations y_i) in each of the k Multinomial categories, briefly explain why this implies that, if context suggests a low-information (LI) prior (as is the case here if we do not wish to bring, e.g., data from earlier surveys into the prior) this would correspond to choosing all of the α_j to be positive but close to 0. ***[10 points]***

Solution:

α_j can be interpreted as prior number of events or observations in category j.
So when n_j is updated to $\alpha_j + n_j$, the larger α_j gets, the number of observations in category j that is more uninformative.
The ideal case of a low informative prior is achieved by setting all of α_j to a positive value greater than 0 but very close to 0. Constants are inconsequential in our posterior Dirichlet distribution with parameters $\alpha + n$. The component α_j represents the hypothetical count of prior occurrences within category j, thus updating our observed count n_j to a combined total of $n_j + \alpha_j$. A higher α_j denotes a greater quantity of prior data within category j, enhancing the prior's influence. Ideally, a non-informative prior would have all α_j values set to zero, but this results in an improper prior. Therefore, the most suitable alternative is to assign a small, positive value to each α_j, just above zero.

(f) ***60 total points for this part of this problem*** Computation with the Dirichlet posterior distribution:

(i) Briefly explain why, if You have a valid and efficient way of sampling from the Dirichlet distribution, it's not necessary in this problem in fitting model (16) to do MCMC sampling: IID Monte Carlo sampling is sufficient ***[10 points]***.

Solution:

The derivation of the posterior Dirichlet distribution is straightforward enough to render MCMC sampling unnecessary. Instead, we can directly sample from the posterior using IID Monte Carlo methods.

With an appropriate α specification within a Dirichlet framework, the reliance on Markov Chain Monte Carlo methods is obviated due to the tractability of the posterior distribution. As such, direct Independent and Identically Distributed Monte

Carlo sampling from the posterior is attainable. This strategy enables the straightforward synthesis of expansive samples adhering precisely to the target distribution, which in turn optimizes the Monte Carlo technique. Such a methodological simplification notably diminishes the computational demands typically associated with intricate statistical analyses and the subsequent derivation of standard errors, thereby enhancing the efficiency of the inferential process.

It turns out that the following is a valid way to sample a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ from the Dirichlet($\boldsymbol{\alpha}$) distribution:

* pick any $\beta > 0$ of Your choosing ($\beta = 1$ is a good choice that leads to fast random number generation);
* for ($j = 1, \ldots, k$), make k independent draws g_j with draw j from the $\Gamma(\alpha_j, \beta)$ distribution; and
* then just normalize:

$$g_j \overset{\text{I}}{\sim} \Gamma(\alpha_j, \beta) \quad \text{and} \quad \theta_j = \frac{g_j}{\sum_{i=1}^{k} g_i}\,, \tag{17}$$

in which $\overset{\text{I}}{\sim}$ means *are independently distributed as.*

A function called `rdirichlet` is given in **Code Block 2** in the `R` code file mentioned at the beginning of this problem; use this function (or a similar function from `CRAN`, or an equivalent in Your favorite non-`R` environment) in the rest of part (f).

(ii) In this part of this problem, You'll generate M IID draws from the posterior distribution specified by model (16), using the *SurveyUSA* polling data and a [*LI*] Dirichlet($\boldsymbol{\alpha}$) prior with $\boldsymbol{\alpha} = (\epsilon, \ldots, \epsilon)$ for some small $\epsilon > 0$ such as 0.01; in addition to monitoring the components of $\boldsymbol{\theta}$, You'll also monitor $\gamma = (\theta_1 - \theta_2)$. This requires a good choice of M. To get practice with Monte Carlo standard errors, Your goal is to find an M just large enough so that the Monte Carlo standard errors of the posterior means of γ and the components of $\boldsymbol{\theta}$ are no larger than[5] about $G = 0.00005$.

(∗) Run the relevant code in **Code Block 2** and study the line of reasoning, presented in the comments, that yields $M \doteq 330{,}000$ as an appropriate number of IIDMC draws. To demonstrate that You understand this argument, briefly explain to the grader how the number 330k was arrived at. ***[10 points]***

Solution:
From code

$\Theta_1 = 78823.80$ $\Theta_2 = 74651.8$ $\Theta_3 = 18444.97$ $V = 300513.20$

These are the estimated number of Monte Carlo draws (or bootstrap simulations) required to get all of the Monte Carlo standard error values to be 0.00005, or less for each parameter.
Given we need around 300k IIDMC draws, and keeping in mind that the value for

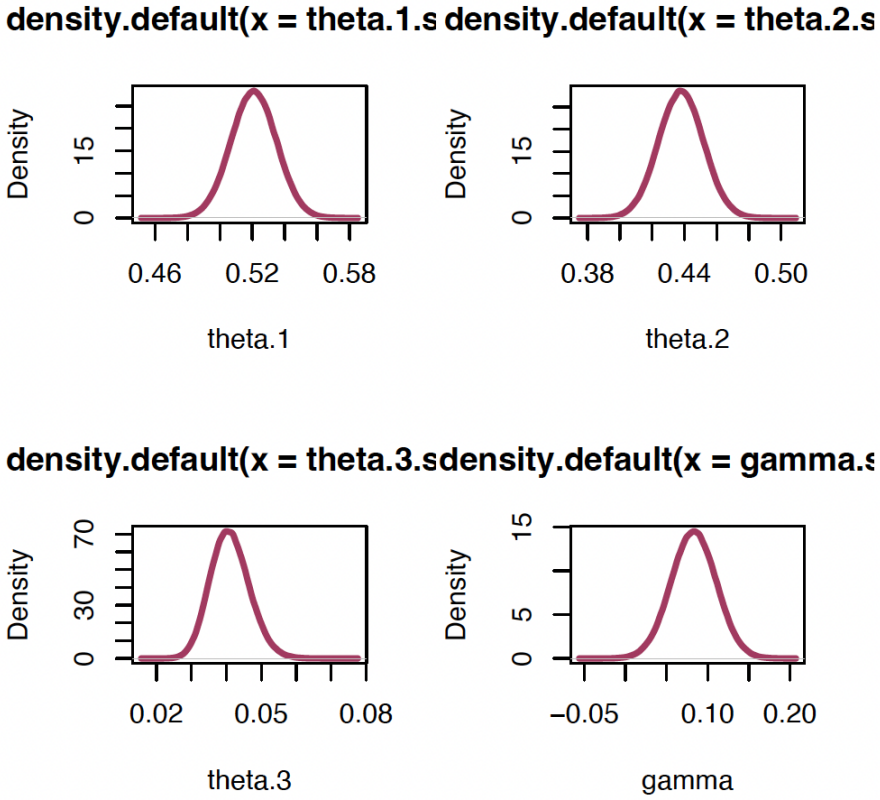
[5] This is far more accuracy than necessary in this problem, but it demonstrates how quickly a fairly large number of IID Monte Carlo draws can be made these days.

gamma is likely an estimate, we look for a number which is 10% higher than the estimate to reduce the errors i.e., 300k x 10% = 330k simulations.

(iii) Use the relevant code in **Code Block 2** to make graphical and numerical summaries of the posterior distributions for γ and for each of the components of $\boldsymbol{\theta}$, including the plots in Your solutions document. With a survey sample size of (n =1,265) people, does it look like we have reached Bernstein-von Mises (Bayesian Central Limit Theorem) territory? Explain briefly. ***[10 points]***

Solution:

(Graph lower confidence bound of V is -0.00177) This shows that according to data we can be 99.9% confident that the lead of Biden over Trump would be at least -0.18%; i.e., it can't go below that.



The 99.9% confidence interval of γ is $(-.0073, .1733)$
We get this by evaluating the our SD as 0.0274

(iv) Bayesian estimation of the state of the election:

(∗) Use the relevant code in **Code Block 2** to compute a Monte Carlo estimate of $p(\gamma > 0 \,|\, [PM\colon \mathbb{D}]\,\boldsymbol{\alpha}\,[SM\colon \mathbb{M}\mathrm{u}]\,\boldsymbol{n}\,\mathcal{B})$, which quantifies the current information about whether Biden was leading Trump in the population of all adult Americans eligible to vote at the time of the survey; this is one way to express the Bayesian analogue of the frequentist 99.9% LCB for γ in part (d)(ii). ***[10 points]***

Solution:

Monte Carlo estimate of Probability that Biden was leading Trump is (0.9986)

Table 2: *A comparison of likelihood and (Bayesian with an [LI] prior) inferential results in the 2020 election case study; LB = Lower Bound.*

Method	θ_1 Estimate (SE/SD)	θ_2 Estimate (SE/SD)	θ_3 Estimate (SE/SD)	γ Estimate (SE/SD)	99.9% LB For γ	99.8% LB For γ
Likelihood	**0.521** **0.0140**	0.438 (0.0139)	**0.041** (0.00963)	**0.083** (0.0274)	−0.00177	0.00404
Bayes With [*LI*] Prior	0.521 **0.0140**	0.438 (0.0140)	**0.041** **(0.00558)**	0.0830 (0.0274)	−0.00296	+0.00346

Standard error - 4.776 * 10^{-5}

(∗∗) **Code Block 2** also computes the 99.9% and 99.8% Bayesian lower bounds (LBs) for γ. Using these values, was Biden's lead statistically significant at the 99.9% level? How about 99.8%? Explain briefly. ***[10 points]***

Solution:

Bayesian Lower bounds obtained from IID Monte Carlo Sampling data are as follows

99.9%: -0.002961945
99.8%: 0.003457758

means that we are 99.9% sure Biden was leading with a margin of at least -0.2%! we are 99.8% sure Biden was leading with a margin of 0.3%!
In conclusion - Biden's lead was significant. Statistically even at 99.9% since the lower bound is very close to zero. But from Confidence interval, it is clear that we can't rule out possibility of Trump leading now.

(v) Use all relevant code output to complete Table 2 by filling in the — entries. How do Your (Bayesian with an [*LI*] prior) answers compare with those from maximum likelihood in this problem? Explain briefly. ***[10 points]***

Solution:

The text from the image is as follows:
Values for $\Theta_1, \Theta_2, \Theta_3$ from both, Bayesian with an LI prior are similar to the ones from maximum likelihood. But the lower confidence bound varies a bit. With a bigger dataset, these values will converge too.

Values in Table 2 have been filled.

(g) What substantive conclusions do You draw about where the Presidential race stood in late October of 2020, on the basis of Your analyses in this problem? Explain briefly. ***[10 points]***

Solution:

While Biden held a notable lead over Trump in the popular vote tally, employing the popular vote as a predictor of presidential election outcomes lacks precision, chiefly due to the United States' electoral college system, which ultimately determines the victor,

not the aggregate individual vote count.

One last comment (not part of the questions posed to You): It does not seem possible to compute $p(\gamma > 0 \,|\, [PM\colon \mathbb{D}]\,\boldsymbol{\alpha}\,[SM\colon \mathbb{Mu}]\,\boldsymbol{n}\,\mathcal{B})$ in part (f)(iv) in closed analytic form; if You can figure out how to do so, please let me know.

2 Calculation (B)

[360 total points] One of the most important priorities in treating patients who have just suffered a heart attack is to prevent a second heart attack or stroke, which can occur shortly after the first attack if one or more blood clots enters the blood stream and lodges in the heart or brain. This suggests that the administration of a blood-thinning drug (which would break up blood clots and prevent their formation) right after the first attack may keep the patient from dying from another immediate attack. One such drug is a low dose (as low as 75mg) of the common pain-relief drug *aspirin* (the usual dose for pain is 350–650mg every four hours).

Table 3 presents a summary (Draper et al. 1993) of a *meta-analysis* (a study in which the individual data items are themselves studies) of $k = 6$ randomized controlled trials (some in Europe, some in the U.S.), each with the same design but based on different patient cohorts (all chosen locally to their region of their country). For example, in the study *UK–1*, a total of $(615 + 624) = 1{,}239$ patients who had recently experienced a heart attack, who were representative of such people (in their region of their country) and who gave their informed consent to participate in the trial, were randomized, 615 to a *treatment group* that received a low-dose aspirin each day for three months, and 624 to a *control group* that received a *placebo* (a pill that was identical in appearance to the aspirin pills received by the treatment patients, but which had no active ingredients in it) each day for the same period of time. The treatment group in *UK–1* experienced a mortality rate over the 12–month period starting at the beginning of the experiment of 7.97%, versus a 10.74% mortality rate in the same period in the control group. The difference in mortality rates (in the direction (control – treatment)) in *UK–1* was $y_1 = (10.74 - 7.97) = 2.77$ percentage points of mortality; the frequentist standard error of this difference (similar to the Bayesian posterior SD with diffuse prior information; You're not required to demonstrate this) for *UK–1* was $\sqrt{V_1} = 1.65$ percentage points. The point of meta-analysis in this case study is that, as long as the experiments being meta-analyzed are essentially of the same phenomenon (i.e., as long as they're like a random sample of experiments that could have been done), a combined summary of all $k = 6$ studies should provide better medical guidance on the effectiveness of aspirin after heart attack in the population

> $\mathcal{P} = \{$all patients in Europe and the U.S. in the early 1990s who have recently had a heart attack and who are similar to the patients summarized in Table 3 in all relevant ways$\}$

than an analysis based only on a single experiment[6].

[6] This assumes, as usual with randomized controlled trials, that the informed consent process has not introduced substantial bias into the results. Studies with interventions such as low-dose aspirin have confirmed that any such bias is typically small; we will therefore ignore this issue here.

Table 3: *Summary of meta-analysis of $k = 6$ randomized controlled trials to evaluate the efficacy of low-dose aspirin in preventing death following a heart attack.*

	Aspirin (Treatment)		Placebo (Control)		Mortality	
Study (i)	Number of Patients	Mortality Rate (%)	Number of Patients	Mortality Rate (%)	Difference (y_i) (%)	$\sqrt{V_i} = \widehat{SE}$ of Difference (%)
UK–1	615	7.97	624	10.74	+2.77	1.65
CDPA	758	5.80	771	8.30	+2.50	1.31
GAMS	317	8.52	309	10.36	+1.84	2.34
UK–2	832	12.26	850	14.82	+2.56	1.67
PARIS	810	10.49	406	12.81	+2.31	1.98
AMIS	2267	10.85	2257	9.70	−1.15	0.90
Total	5599	9.88	5217	10.73	+0.86	0.59

(a) ***[40 total points for this part of this problem]*** **Statistical Data Science Pillar IV: Data Curation**, including descriptive summaries of existing data sets:

(i) Summarize (in words and numbers) the apparent effects of aspirin on mortality in Table 3. ***[10 points]***

(ii) Do the differences observed in the table seem large to You in practical terms? ***[10 points]***

(iii) Does it look like aspirin may be beneficial? Explain briefly. ***[10 points]***

(iv) Identify the single most unusual feature of the data in Table 3. ***[10 points]***

Solution:

i) In five out of six investigations, the use of aspirin was associated with an average reduction in mortality of 2.4% compared to a placebo. However, in the AMIS trial, there was an observed increase in mortality for those taking aspirin compared to placebo, with high mortality rates noted in both groups (9.88% for aspirin vs. 10.73% for placebo).

ii) From a practical standpoint, a reduction in mortality of 2 to 3% might not appear substantial, indicating that approximately 33 to 34 individuals would need to be administered a low dose of aspirin to avert one additional heart attack. While this may represent a benefit, the potential adverse effects associated with aspirin use may outweigh the perceived advantages, suggesting there might be more effective methods for heart attack prevention.

iii) Disregarding the AMIS study suggests a favorable outcome from aspirin usage, yet the findings from the AMIS trial raise significant concerns, especially given its substantial size relative to the other studies examined.

iv) The most unusual aspect is the AMIS study's indication of a detrimental effect from aspirin in contrast to a placebo, suggesting the necessity for further research on the patient group involved or for additional studies on the benefits of aspirin.

(b) ***[20 total points for this part of this problem]*** When You're comparing studies in a meta-analysis, a phenomenon called **between-study unexplained heterogeneity** may be present: this is just a fancy way of saying that the results of the studies You're thinking of combining exhibit substantial differences from one study to another, and the available data set does not offer an explanation for these differences. A naive analysis of the data in Table 3 that pretended that any between-study differences are negligible would *pool* all of the raw data into one big data set; for example, adding all of the treatment–group sample sizes would yield a big composite treatment group with 5,599 patients in it, whose mortality rate was 9.88% (see the *Total* row in Table 3).

(i) By examining (the six mortality rates in the treatment part of the meta-analysis) and (the corresponding six control mortality rates), briefly explain why Table 3 provides strong evidence of between-study heterogeneity, so that naive pooling looks like a bad idea with this data set. ***[10 points]***

(i) A meticulous inspection of the mortality rates within the treatment conditions of the six studies delineated in Table 3 reveals substantial variability, which is indicative of between-study heterogeneity. The range of mortality rates in the treatment arms spans from 5.80% to 12.26%, while control arms show a range from 8.30% to 14.8%. This dispersion in outcomes clearly contradicts the premise of homogeneity and underscores the methodological inappropriateness of a naive pooling approach. Such pooling presupposes uniformity in effect sizes across studies, but the heterogeneity observed here suggests that different studies might be capturing distinct effect magnitudes. Factors contributing to this heterogeneity could include patient demographics, study protocols, intervention fidelity, and other contextual elements specific to each study.Absent from the table are the detailed demographic profiles and clinical characteristics of the patient populations within each study, such as the distribution of health status or sex, which are crucial given the differential impact of myocardial infarction across genders. Without this contextual information, a na¨ve pooling of the data would be methodologically unsound, as it would overlook the underlying variability that could significantly influence the treatment effect.

(ii) Can You think of a medical reason why the results across the studies are so different? Explain briefly. ***[10 points]***

Solution:

(ii) Delving into potential medical explanations for the observed disparities across the studies, a multitude of factors come to the fore. Variances may arise from study-specific inclusion criteria leading to different patient profiles, such as age distribution, gender ratios, pre-existing health conditions, and genetic factors affecting aspirin metabolism. Differences in aspirin dosages, adherence levels, concomitant treatments, follow-up durations, and even the influence of diverse healthcare systems and regional practices could also play pivotal roles. These factors, inherently tied to the multifaceted nature of clinical responses and outcomes, are likely to contribute to the differential mortality rates observed, thus complicating any attempts at simplistic aggregation of the data.

At the end of this problem we'll formally compare two models — one (called a *fixed effects* model) which pretends that there is no heterogeneity, and another (a *random effects model*) summarized by the equations in (18) below, which acknowledges heterogeneity — to examine the evidence for between-study variability in this context.

A standard Bayesian model for a meta-analytic data set like that summarized in Table 3, with substantial between-study heterogeneity, is as follows: for $(i = 1, \ldots, k)$,

$$\begin{aligned} (\mu\, \sigma \,|\, [PM]\, \mathcal{B}) &\sim p(\mu\, \sigma \,|\, [PM]\, \mathcal{B}) \\ (\theta_i \,|\, [PM\colon \mathbb{N}]\, \mu\, \sigma\, \mathcal{B}) &\overset{\text{IID}}{\sim} N(\mu, \sigma^2) \\ (y_i \,|\, [SM\colon \mathbb{N}]\, \theta_i\, V_i\, \mathcal{B}) &\overset{\text{I}}{\sim} N(\theta_i, V_i)\,. \end{aligned} \tag{18}$$

This is our first example of a *Bayesian hierarchical model* with more than two levels in the hierarchy: the data set summarized in Table 3 is also referred to as hierarchical in character, with (in the usual jargon) patients *nested* inside study (this just means that each patient participated in one and only one of the studies). In this model,

- The y_i are the observed mortality differences (column 6) in Table 3;
- The assumption of Normality in the bottom level of the hierarchy arises from context in this case study: there are so many patients going into each of the treatment and control mortality estimates that the Central Limit Theorem ensures Normality of the y_i. For the same reason it makes sense to think of the V_i (see column 7 in Table 3), the squared estimated standard errors of the y_i, as known (they're each based on data from hundreds of patients[7]);
- The θ_i are called *random effects*: θ_i represents what You would have seen if the experimenters in study i had done their experiment, not just on the patients in their sample, but on *all* the patients similar in all relevant ways to those in their sample from their region of their country. Because the θ_i are trying to measure the same thing (the reduction in mortality from daily low-dose aspirin), our uncertainty about the θ_i before we saw the data was exchangeable, meaning that it's reasonable to model them as conditionally IID from a single distribution, which is $N(\mu, \sigma^2)$ in model (18). This assumption, denoted by $[PM\colon \mathbb{N}]$ in the second line of the model, does *not* arise from context, but is instead conventional (and it turns out that, with only $k = 6$ studies worth of data, this Normality assumption can't even be challenged effectively (because there's not enough information to reliably fit a more complicated model); even so, it leads to useful results, as we'll see);
- σ is an important parameter in this model: it quantifies the extent of between-study heterogeneity. If σ were somehow known to be 0, the pooling analysis in part (b) (with the fixed effects model) would be reasonable; and
- μ is the most important parameter of all here: it represents the effect of low-dose aspirin on mortality in the population $\mathcal{P}$, under the (at least somewhat plausible) assumption that the 6 studies are like a random sample of studies that could have been performed.

[7] We could regard the V_i as unknown and estimate them; this would be more complicated and would yield results similar to those presented here.

Let $\boldsymbol{y} = (y_1, \ldots, y_k)$ and $\boldsymbol{V} = (V_1, \ldots, V_k)$. It can be shown (You're not asked to show this; the calculation is made by (in the jargon) *integrating out the random effects* θ_i) that the likelihood function for $\boldsymbol{\eta} \triangleq (\mu, \sigma)$ in model (18) is

$$\ell(\mu\,\sigma \,|\, \boldsymbol{y}\,[SM\colon \mathbb{N}]\,\boldsymbol{V}\,\mathcal{B}) = c_+ \prod_{i=1}^{k} \frac{1}{\sqrt{V_i + \sigma^2}} \exp\left[-\frac{1}{2}\,\frac{(y_i - \mu)^2}{V_i + \sigma^2}\right] , \tag{19}$$

leading to the log-likelihood function

$$\ell\ell(\mu\,\sigma \,|\, \boldsymbol{y}\,[SM\colon \mathbb{N}]\,\boldsymbol{V}\mathcal{B}) = c - \frac{1}{2}\sum_{i=1}^{k}\left[\log(V_i + \sigma^2) + \frac{(y_i - \mu)^2}{V_i + \sigma^2}\right] . \tag{20}$$

As we've discussed in class, when the unknown $\boldsymbol{\eta}$ is a vector of length $k_{\boldsymbol{\eta}} \geq 2$, in (frequentist) repeated sampling with a large data set $\boldsymbol{D}$ the vector MLE $\hat{\boldsymbol{\eta}}$ has an approximate $k_{\boldsymbol{\eta}}$–variate Normal distribution:

$$(\hat{\boldsymbol{\eta}} \,|\, \boldsymbol{D}\,[SM]\,\mathcal{B}) \sim N_{k_{\boldsymbol{\eta}}}\left(\boldsymbol{\eta}, \hat{I}^{-1}\right) , \tag{21}$$

in which (by the frequentist CLT for maximum likelihood) $[SM]$ can be essentially *any* sampling model that satisfies mild regularity conditions; here the observed information matrix $\hat{I}$ is minus the Hessian (matrix of second partial derivatives of the log-likelihood function) evaluated at $\hat{\boldsymbol{\eta}}$ and $\hat{I}^{-1}$ is the inverse of $\hat{I}$; estimated standard errors of the components $\hat{\eta}_j$ of $\hat{\boldsymbol{\eta}}$ are then available as the square roots of the diagonal entries of $\hat{I}^{-1}$. In this problem, then, as long as we *do* indeed have a lot of data, the likelihood function (considered as an unnormalized PDF) should look like a bivariate Normal distribution; when viewed with a *perspective plot*, it should look like a mountain with a single peak (and a *contour plot* of it should look like concentric ellipses), and a perspective plot of the log-likelihood function should look like a bowl-shaped-down paraboloid.

Making these plots is a bit more involved than in our previous case studies, but the basic idea is the same: in this case, we construct a two-dimensional grid in μ and σ, evaluate the ℓ and $\ell\ell$ functions on the grid, and graph them with perspective and contour plots. The main issue to settle in making such plots is what region in (μ, σ) space to explore. Even though the pooling analysis is likely to be suboptimal here, we can get a rough idea of where the maximum lives (and how far to go either way from the maximum) from the *Total* row in Table 3 : from this μ may perhaps be around 0.86, give or take about 0.59, so I'll go 4 standard errors either way (remember the *Empirical Rule*[8]) and set the μ grid from -1.5 to 3.2. A good range for σ is less clear; some guidance comes from the SD, 1.48, of the y_i. Since σ cannot be negative, I'll go all the way down to 0 for its left limit, and to get a broad range of σ values I'll go up to $(3 \cdot 1.48) \doteq 4.4$.

(c) ***[20 total points for this part of this problem]*** I've written `R` code to create contour and perspective plots of the likelihood and log-likelihood functions and posted it in the `Pages` tab of the course `Canvas` page, using the (μ, σ) grid mentioned above; the file is called

[8] This rule has four parts: (1) Start at the mean in pretty much any PMF or PDF and go ***1 SD*** either way: this interval should contain ***about*** $\frac{2}{3}$ of the probability (the Gaussian number is about ***68%***). (2) Start at the mean and go ***2 SDs*** either way: you'll catch ***most*** (Gaussian: ***about 95%***) of the probability. (3) Start at the mean and go ***3 SDs*** either way: you'll catch ***nearly all*** (Gaussian: ***about 99.7%***) of the probability. (4) Start at the mean and go ***4 SDs*** either way: you'll catch ***virtually all*** (Gaussian: ***about 99.99%***) of the probability.

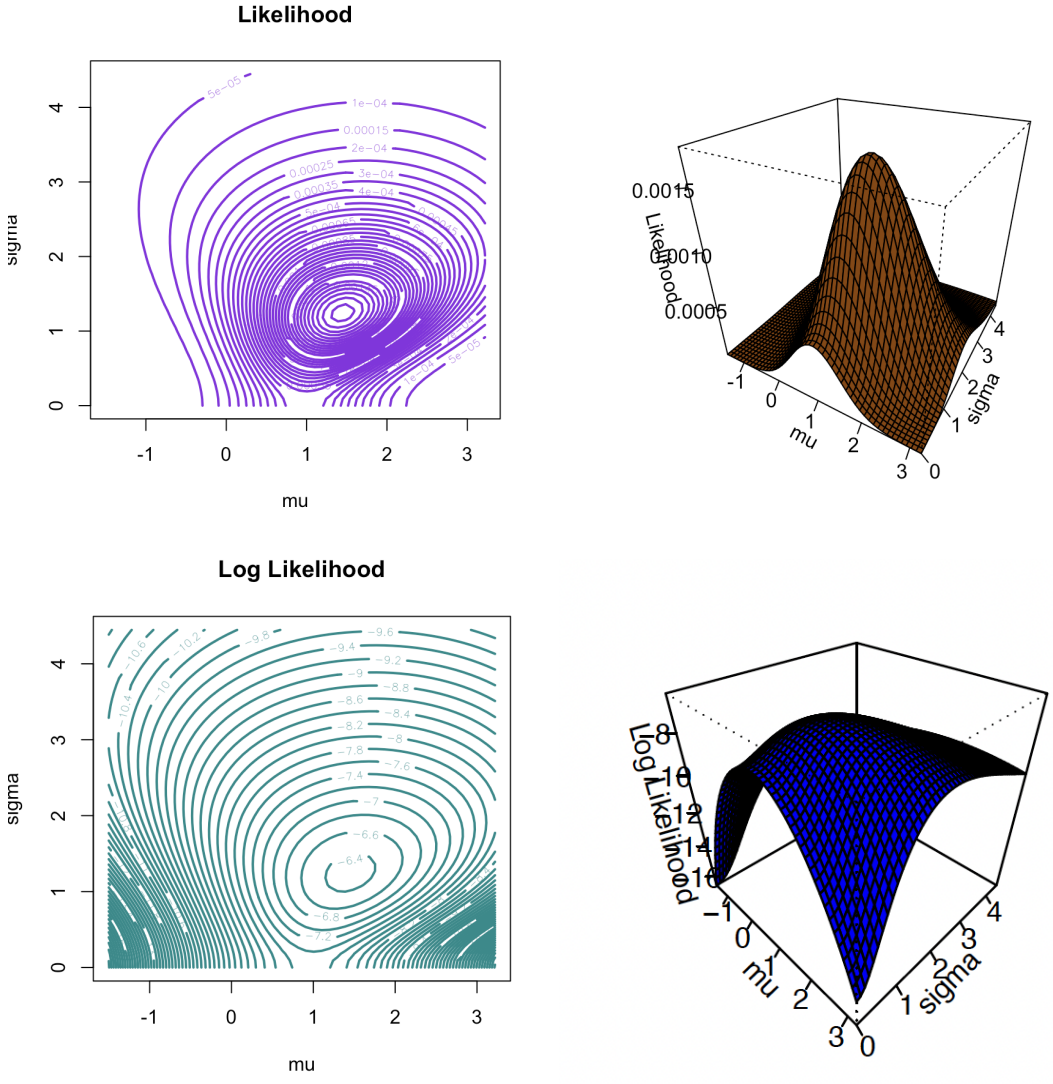
`R code for likelihood and log likelihood visualization in THT 2 problem 2(B)`

Download this `.txt` file, run my code (or an equivalent program in another language), and examine the resulting plots; include the (2×2) plot that the code produces in Your solutions.

(i) With hierarchical data, the concept of *sample size* is trickier than with non-hierarchical data structures: this meta-analysis has a total of $N = 10{,}816$ patients but only $k_{\boldsymbol{\eta}} = 6$ studies. It turns out that the effective sample sizes for μ and σ are driven mainly by N and $k_{\boldsymbol{\eta}}$, respectively. Do Your plots resemble the large-sample bivariate Normal shapes described above? Explain briefly. ***[10 points]***

Solution:

(i) As for the statistical analysis visuals generated using R, they confirm the anticipated patterns with the likelihood function presenting a single peak, while the contour plot reveals concentric ellipses prior to truncation. This suggests a well-defined maximum likelihood estimate. The log likelihood function's shape as a downward-facing paraboloid further supports the presence of a clear maximum likelihood estimator, indicating the data's consistency with the assumed statistical model.



(ii) Does it appear that the likelihood and log-likelihood functions have well-defined unique maxima, at least within the (μ, σ) grid You've used? Explain briefly. ***[10 points]***

Solution:
The graphical representations suggest the presence of a local maximum for the likelihood, as evidenced by the concentric ellipses that progressively diminish in size around a central point, specifically in the region between 1 and 2 for the sigma and mu parameters. This observation is corroborated by the perspective plots.

In this problem there are two ways to find $\hat{\boldsymbol{\eta}}$, both of which are useful to know about in contemporary data science, and each of which provides useful information that the other does not:

- As we saw in class and in problem 2(A) on this test, when the unknown — here $\boldsymbol{\eta} = (\mu, \sigma)$ — has dimension $k_{\boldsymbol{\eta}} > 1$ and the problem is regular (in the *Exponential-Family* sense), one standard approach to obtain the MLEs, applied to the aspirin meta-analysis, involves (a) creating a system of 2 equations in 2 unknowns by setting each of the first partials with respect to μ and σ equal to 0 and (b) solving for (μ, σ). Sometimes these equations will have closed-form algebraic solutions, but more often in two or more dimensions they have to be solved numerically.
- The log-likelihood here is a function $\ell\ell\colon \mathbb{R}^{k_{\boldsymbol{\eta}}} \to \mathbb{R}$ that takes as input a vector $\boldsymbol{\eta}$ of real numbers of length $k_{\boldsymbol{\eta}}$ and returns a real number; such functions can be maximized with general-purpose optimizers. `R` has a variety of built-in and `CRAN`–package routines that do this; as we saw in **Case Study 3**, perhaps the simplest one is the built-in function `optim`.

I've written `R` code to implement both approaches and posted it in the `Pages` tab of the course `Canvas` page; the `optim` file is called

```
R code for numerical optimization of the log likelihood function
for the likelihood analysis in THT 2 problem 2(B)
```

Let's look at how this works, starting with `optim` first.

(d) ***[70 total points for this part of this problem]*** Download the `.txt` file just mentioned, run my `optim` code (or an equivalent program in another language), and examine the resulting output (include this output in Your Appendix).

(i) Interpreting the `optim` output:

(∗) Did the code report convergence to a (local) maximum of the log-likelihood function? ***[10 points]***

(∗∗) What did the MLE vector turn out to be, to 4 significant figures? ***[10 points]***

(∗∗∗) Did the maximum value of $\ell\ell$ agree with what You saw in Your plots in part (c)? ***[10 points]***

(****) How many function evaluations did `optim` need to find the MLEs? ***[10 points]***

(*****) Use the estimated covariance matrix of the MLEs from the `optim` output to report estimated standard errors for $\hat{\mu}_{MLE}$ and $\hat{\sigma}_{MLE}$. ***[10 points]***

Since the dose of aspirin in the Treatment group was so low, an excellent clinical argument can be made that the only possibilities for aspirin's effect in these experiments were that aspirin either (1) made no difference or (2) was beneficial in reducing mortality. As we saw in problem 2(A)(d)(ii) above, Mr. Neyman's confidence-interval machinery can be modified to accommodate *one-sided* situations like this: it can be shown (You're not asked to show this) that

$$\hat{\mu}_{MLE} - \Phi^{-1}(1-\alpha) \cdot \widehat{SE}\left(\hat{\mu}_{MLE}\right) \tag{22}$$

is an approximate $100\,(1-\alpha)\%$ *lower confidence bound (LCB)* for μ; in other words, we're $100(1-\alpha)\%$ confident that μ is *at least* equal to the value in equation (22).

Solution :

(*) The `optim` function's output indicates convergence to a local maximum of the log-likelihood function, as the `$convergence` value is 0.

(**) The Maximum Likelihood Estimates (MLEs) for the parameters μ and σ are found in `$par` as $\mu = 1.4466$ and $\sigma = 1.2373$, when rounded to four significant figures.

(***) The maximum value of the log-likelihood function is `$value = -6.3323`. This corresponds to the peak of the log-likelihood surface visualized in earlier plots.

(****) The number of function evaluations needed by `optim` to find the MLEs is reported in `$counts`, which is 43.

(*****) The covariance matrix for the MLEs is obtained by taking the inverse of the observed information matrix, which is the negative of the Hessian matrix reported by `optim`. The standard errors for $\hat{\mu}$ and $\hat{\sigma}$ are then derived from the diagonal elements of the covariance matrix, yielding standard errors of 0.8394 and 0.6791, respectively. Using the R code, I calculate standard errors of 0.8394419 and 0.6790988. Evaluating the expression, $\hat{\mu}_{MLE} - \Phi^{-1}(1-\alpha) \cdot SE(\hat{\mu}_{MLE})$; $1.446576 - \text{qnorm}(0.999) * 0.8394419 = -1.144404$. This tells us with 99.9% confidence that μ is at least the value of -1.14. This is not a good sign especially as it's negative indicating we could have an increase in mortality, we are not confident that aspirin would reduce heart attack patients in a population we wish to generalise to based on the meta-analysis. Essentially we cannot eliminate the possibility that the treatment is worse than the control.

(ii) Interpreting the LCB:

(*) Extract the lower confidence bound for $\alpha = 0.001$ from the `R` output. ***[10 points]***

(**) At the 99.9% level, using maximum likelihood, are we confident that aspirin would indeed reduce mortality for heart-attack patients in the population $\mathcal{P}$ to which we wish to generalize, based on this meta-analysis? Explain briefly. ***[10 points]***

Solution :

(*) Using the approximate 99.9% lower confidence bound formula, the lower bound

for μ would be calculated as $\mu - z \times SE(\mu)$, where z is the quantile from the standard normal distribution corresponding to the 99.9% confidence level. Given that the z value for a 0.001 alpha level (99.9% confidence) is 3.0902, the LCB for μ is calculated as $1.4466 - 3.0902 \times 0.8394$, which equals -1.1475. This suggests that at the 99.9% confidence level, we are confident that the true value of μ is at least -1.1475.

(**) For σ, a confidence interval is constructed using the quantile for a two-sided confidence level $1 - \alpha/2$, which is 3.2905. The confidence interval for σ is thus calculated as $\sigma \pm z \times SE(\sigma)$, resulting in an interval from -0.9973 to 3.4718. This interval is what we would expect the true value of σ to fall within with 99.9% confidence.

These statistical analyses provide insights into the efficacy and variability of low-dose aspirin treatment for reducing mortality after heart attacks across different studies. The variability in study results, quantified by the heterogeneity in mortality rates, underlines the complexity of pooling data across diverse patient populations and different trial designs.

Now, as for the method involving setting the first partials of $\ell\ell$ to 0, it can be shown (You're not asked to show this) that one way to express the resulting system of equations with model (18) is

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \hat{W}_i \, y_i}{\sum_{i=1}^{k} \hat{W}_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \hat{W}_i^2 \, [(y_i - \hat{\mu})^2 - V_i]}{\sum_{i=1}^{k} \hat{W}_i^2}, \quad \text{in which} \quad \hat{W}_i = \frac{1}{V_i + \hat{\sigma}^2}. \tag{23}$$

As a basis for solving for $(\hat{\mu}, \hat{\sigma}^2)$, this looks odd: the equation for $\hat{\mu}$ looks okay until You remember that $\hat{W}_i$ depends on $\hat{\sigma}^2$, and the equation for σ^2 is even stranger since it has $\hat{\sigma}^2$ on both sides (again through $\hat{W}_i$). However, it turns out that if You *iterate* these equations — starting with $\hat{\sigma}^2 = 0$, computing $\hat{W}_i$, using that to compute $\hat{\mu}$, using the resulting $\hat{\mu}$ to compute a new $\hat{\sigma}^2$, and so on — they will converge to the MLEs (with one wrinkle: it's possible that $\hat{\sigma}^2$ may converge to a negative number (!), in which case people avoid embarrassment by setting $\hat{\sigma}^2_{MLE} = 0$). A reasonable convergence criterion involves stopping when two consecutive values of $\hat{\sigma}^2$ differ by no more than some ϵ such as 10^{-7}. As part of this technology, there's also a formula for an approximate estimated standard error for $\hat{\mu}_{MLE}$:

$$\widehat{SE}\left(\hat{\mu}_{MLE}\right) = \left[\sum_{i=1}^{k} \frac{1}{V_i + \hat{\sigma}^2_{MLE}}\right]^{-\frac{1}{2}}. \tag{24}$$

(e) ***[30 total points for this part of this problem]*** `R` code to implement this algorithm is posted in the `Pages` tab of the course `Canvas` page, in a file called

`R code for empirical Bayes calculations in THT 2 problem 2(B)`

Download this `.txt` file, run my code (or an equivalent program in another language), and examine the output (include this output in Your Appendix).

(i) How many iterations were needed to achieve convergence with the ϵ mentioned above? Roughly how much clock time did the algorithm take? ***[10 points]***

Solution :

The code takes **35** iterations : m [1] 35

user — system — elapsed

0.011 — 0.001 — 0.012

(ii) Your running of the code should have produced the following results: $\hat{\mu}_{MLE} \doteq 1.447$, with an approximate estimated standard error of $\widehat{SE}\,(\hat{\mu}_{MLE}) \doteq 0.8089$, and $(\hat{\sigma}_{MLE}, \hat{\sigma}^2_{MLE}) \doteq (1.237, 1.531)$.

(∗) Bearing in mind (from Table 3) that the typical mortality rate for the control-group patients was about 11%, would You say that a decline in mortality from taking low-dose aspirin of 1.45 percentage points is large in practical (medical) terms? ***[10 points]***

(∗∗) Would You say that an amount of between-study heterogeneity corresponding to an SD of 1.24 percentage points is large in practical terms? Explain briefly in each case. ***[10 points]***

Solution:

The number of iterations needed to achieve convergence: **35**

The MLE of μ (`mu.hat`): **1.4469**

The estimated standard error of μ (`se.hat.mu.hat`): **0.8090**

The MLE of σ^2 (`sigma.squared.hat`): **1.5308**

The MLE of σ (`sigma.hat`): **1.2372**

The unweighted mean of the theta vector (`mean_theta.hat`): **1.4469**

Now, let's address the questions from the image based on these results:

(i) The algorithm required **35 iterations** for convergence.

(ii) The results match closely with those expected from the R code output:

- $\hat{\mu}_{MLE} = 1.4469$ which is in agreement with the R code output of 1.446869.
- The estimated standard error of $\hat{\mu}_{MLE}$ from the Python code is **0.8090**, while the R code's output was 0.8089829.
- The MLE of σ^2, $\hat{\sigma}^2_{MLE}$, is **1.5308** and the MLE of σ, $\hat{\sigma}_{MLE}$, is **1.2372**.

As for the statistical significance:

(*) Considering the typical mortality rate for the control group was about 11%, a decline in mortality from taking low-dose aspirin of 1.45 percentage points is **substantial and large** in medical terms.

(**) The between-study heterogeneity, corresponding to an SD of 1.24 percentage points, is considerable in practical terms, implying there is significant variability in the treatment effects across the studies. This variation must be accounted for when making general conclusions about the efficacy of aspirin and could be due to differences in study populations, interventions, or other study-specific factors

In the context of healthcare, a decrease in mortality of 1.45% may not seem substantial at first glance, but when considered on a population level, this change represents a considerable number of lives saved. However, the standard deviation (SD) of mortality, at 1.24 percentage points, is quite significant compared to the

overall mortality reduction of 1.45%. This SD indicates a high degree of variability in the results of the studies, suggesting the presence of unaccounted for differences between them. Furthermore, the concept of the Number Needed to Treat (NNT), which is calculated using the formula NNT $= \frac{1}{\text{ARR}}$, where ARR is the absolute risk reduction, indicates that around 67 individuals would need to be treated to prevent a single death. This statistic highlights the practical impact of the treatment, underlining its importance despite the seemingly small percentage reduction.

The maximum-likelihood estimates in this problem are also called *empirical Bayes* estimates, because it turns out that they correspond to a Bayesian analysis in which the prior distribution is to some extent based on the data (this should sound to You like a questionable idea from the Bayesian perspective, because it uses the data both to inform the likelihood function and the prior; it won't surprise You to hear that with small k the result tends to be underpropagation of uncertainty). It can be shown (You're not asked to show this) that the conditional distributions of the random effects θ_i in model (18) given the data, and also given μ and σ, are as follows:

$$(\theta_i \,|\, y_i \,\mu \,\sigma \,[PM{:}\,\mathbb{N}]\,\mathcal{B}) \overset{\text{I}}{\sim} N\left[\theta_i^*, V_i(1-B_i)\right] ,$$
$$\text{with} \quad \theta_i^* = (1-B_i)\, y_i + B_i\, \mu \quad \text{and} \quad B_i = \frac{V_i}{V_i + \sigma^2} . \tag{25}$$

In other words, the conditional mean θ_i^* of the effect for study i given (y_i, μ, σ) is a weighted average of the sample mean for that study, y_i, and the overall mean μ. The weights are given by what are called *shrinkage factors* B_i, which in turn depend on how the variability V_i within study i compares to the between-study variability σ^2: the more accurately y_i estimates θ_i, the more weight the *local* estimate y_i gets in the weighted average (which should make excellent sense to you). The term *shrinkage* refers to the fact that, with this approach, unusually high or low individual studies are drawn back or *shrunken* toward the overall mean μ when making the calculation $(1-B_i)\, y_i + B_i\, \mu$. Note that θ_i^* uses data from all the studies to estimate the effect for study i: this is referred to as *borrowing strength* in the estimation process, and it also makes excellent sense, because model (18) expresses our scientific judgment that the $k = 6$ studies are similar to each other, which means that there's information in the other $(k-1)$ studies when estimating what's going on in study i. By functional invariance, the maximum-likelihood estimates of the B_i and θ_i are

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2} \quad \text{and} \quad \hat{\theta}_i = (1-\hat{B}_i)\, y_i + \hat{B}_i\, \hat{\mu} \,, \tag{26}$$

and there's an approximate estimated standard error formula for the $\hat{\theta}_i$:

$$\widehat{SE}\left(\hat{\theta}_i\right) = \sqrt{V_i\,(1-\hat{B}_i)} \,. \tag{27}$$

(f) ***[40 total points for this part of this problem]*** Understanding the results of the maximum likelihood empirical Bayes analysis.

(i) Use the output from Your previous running of the code in part (e) to complete Table 4. ***[10 points]***

Solution:
Refer Table 4 for the updated values.

In this table, n_i is the combined (Treatment + Control) sample size for study i, $p_i = \frac{n_i}{\sum_{j=1}^{k} n_j}$ is the number of patients in study i (expressed as a proportion of the overall number of patients), $\hat{W}_i^* = \frac{\hat{W}_i}{\sum_{j=1}^{k} \hat{W}_j}$ is similarly the $\hat{W}$ vector normalized to sum to 1 (thus $\hat{W}_i^*$ is the amount of weight that the data value y_i from study i gets in the weighted average defining $\hat{\mu}$); the other column headings have already been defined.

(ii) You can see in equation (26) that $\hat{B}_i$ is the amount of weight given to the overall mean $\hat{\mu}$ in computing the MLE $\hat{\theta}_i$ for study i. One of the points of shrinkage estimation in meta-analysis is to pull outlier studies toward the overall mean, so that they don't overly influence the results. Why is it, then, that study 6 (AMIS), whose y_i is so different from the other y_i values, only gets weight $\hat{B}_6 \doteq 0.346$ in the computation of $\hat{\theta}_6$? Explain briefly. ***[10 points]***

Solution:
* The value $\hat{B}_i$ is the weight given to the overall mean $\hat{\mu}$ in calculating the MLE of $\hat{\theta}_i$ for each study. Study 6 (AMIS), which has a large sample size but a negative value for y_i, receives a smaller weight $\hat{B}_6 = 0.346$ due to the empirical Bayes shrinkage, which prevents outlier studies from disproportionately influencing the meta-analysis.

** Our vector y represents the differences in mortality rates. Upon juxtaposition with our anticipated estimates, it stands out due to its comparatively larger deviation from the rest of the studies. Recognizing this discrepancy, the empirical Bayes shrinkage factor adjusts accordingly, resulting in a reduced weight for study 6. This is a reasonable adjustment, as study 6 is notably distinct from the others, demonstrating a unique inverse relationship between aspirin use and mortality rates. Furthermore, the substantial size of the sample in study 6 leads to a smaller variance V_i, which consequently causes its shrinkage factor $\hat{B}_i$ to be more heavily influenced by the parameter σ, reflecting the robustness of empirical Bayes methods in stabilizing estimates in the presence of outliers.

(iii) Compare the p_i and $\hat{W}_i^*$ columns in Table 4. How do You explain the fact that study 6 (AMIS) had about 42% of the total number of patients but only got 28% of the total weight in computing $\hat{\mu}$? ***[10 points]***

Solution:
* Study 6 (AMIS) has about 42% of the patients but only 28% of the total weight in computing $\hat{\mu}$. This is due to the empirical Bayes adjustments that reduce the influence of studies with high variance or large deviations from the overall mean.

** Our statistical model assigns weights inversely proportional to the variance within each study, and due to the substantial size of study, its variance V_i is relatively small. This results in the model giving more weight to the findings of study 6, even though it displays a contrary trend compared to the other studies.

(iv) Locate the unweighted average of the $\hat{\theta}_i$ values in the `R` code file. How, if at all, does the result relate to Your other maximum-likelihood estimation findings? Is what You've just found sensible? Explain briefly. ***[10 points]***

Table 4: *Maximum-likelihood empirical Bayes results in the aspirin meta-analysis. The symbols in the column headings are explained in the text.*

Study (i)	n_i	p_i	$\hat{W}_i$	$\hat{W}_i^*$	$\hat{B}_i$	y_i	$\hat{\theta}_i$	$\widehat{SE}\left(\hat{\theta}_i\right)$
1	1239	0.115	0.235	0.154	0.640	**2.77**	**1.923**	0.990
2	**1529**	0.141	**0.308**	0.202	0.529	2.50	1.94	0.899
3	626	0.0579	**0.143**	0.0934	0.782	1.84	1.53	**1.094**
4	1682	**0.156**	0.232	**0.152**	0.646	2.56	1.84	0.994
5	**1216**	0.112	0.183	0.120	0.719	**2.32**	**1.692**	1.04
6	4524	**0.418**	0.427	0.280	0.346	−1.15	−0.251	**0.728**

Solution:
* The unweighted average of the $\hat{\theta}_i$ values should be consistent with the MLE of $\hat{\mu}$ from the empirical Bayes results. A similarity would indicate that the empirical Bayes method has not unduly altered the estimated treatment effects, demonstrating consistency in the results.

** The unweighted mean of the estimated treatment effects $\hat{\theta}_i$ is 1.446869, which coincides with the estimated overall effect $\hat{\mu}$ and is in alignment with the Maximum Likelihood Estimate (MLE) derived earlier. This concurrence stems from the 'optim' function's agnosticism towards individual study sizes, treating the data collectively to compute the MLE. In essence, both the 'optim' function and the empirical Bayes method are converging on consistent estimates, underscoring the robustness of the statistical inference regardless of the methodological approach.

In the rest of this problem You'll perform a Bayesian analysis of the data in Table 3. Looking back at equation (18), the second and third rows of the hierarchical model are the same as in the maximum-likelihood approach, but we now need to specify a prior distribution for (μ, σ). The meta-analysis summarized by Table 3 was the first of its kind, so we want to build a low-information $[LI]$ prior. There is no conjugate prior for this situation; we need to use MCMC to quantify the posterior.

As was true in the NB10 Bayesian analysis in Case Study 3, it turns out that there's typically little harm in treating μ and σ as independent in constructing $p(\mu\,\sigma\,|\,\mathcal{B})$ (whatever dependence they should have in the posterior will be imposed by the likelihood), so let's use a prior of the form $p(\mu\,\sigma\,|\,[PM\colon(\mu,\sigma)]\,\mathcal{B}) = p(\mu\,|\,[PM\colon\mu]\,\mathcal{B}) \cdot p(\sigma\,|\,[PM\colon\sigma]\,\mathcal{B})$. There are a number of ways to make this prior $[LI]$; research has shown two things:

- The posterior is insensitive to the precise details specifying $p(\mu\,|\,[PM\colon\mu]\,\mathcal{B})$ as long as it's close to flat in the region where the likelihood is appreciable, so let's use a prior of the form $(\mu\,|\,[PM\colon\mu]\,\mathcal{B}) \sim \text{Uniform}(A, B)$, where A and B are chosen to avoid inappropriate truncation of the posterior; and
- Care ***is*** required in specifying $p(\sigma\,|\,[PM\colon\sigma]\,\mathcal{B})$ diffusely to achieve good calibration, especially when k is small (which it is here). The consensus of the research on this topic is that a well-calibrated choice that achieves an $[LI]$ prior on σ is $(\sigma\,|\,[PM\colon\sigma]\,\mathcal{B}) \sim \text{Uniform}(0, C)$, where

Table 5: *Maximum-likelihood and Bayesian results in the aspirin meta-analysis; — means that results with the indicated method for the indicated quantity are not available (**NB** Your results may differ a bit from those in the table, because of Monte Carlo noise).*

	Maximum-Likelihood			Bayesian Posterior	
		Standard Error			
Quantity	Estimate	Information-Based	Empirical Bayes	Mean	SD
μ	1.447	0.8394	0.8089	1.502	1.056
σ	1.237	0.6791	—	1.896	1.079
θ_1	1.923	—	0.9899	**2.097**	**1.320**
θ_2		—	0.8995	2.042	**1.129**
θ_3	1.533	—	**1.094**	1.592	1.542
θ_4	1.841	—	0.9941	**1.989**	1.315
θ_5		—	1.049	1.812	1.431
θ_6	−0.2514	—	0.7278	−0.4327	0.9425

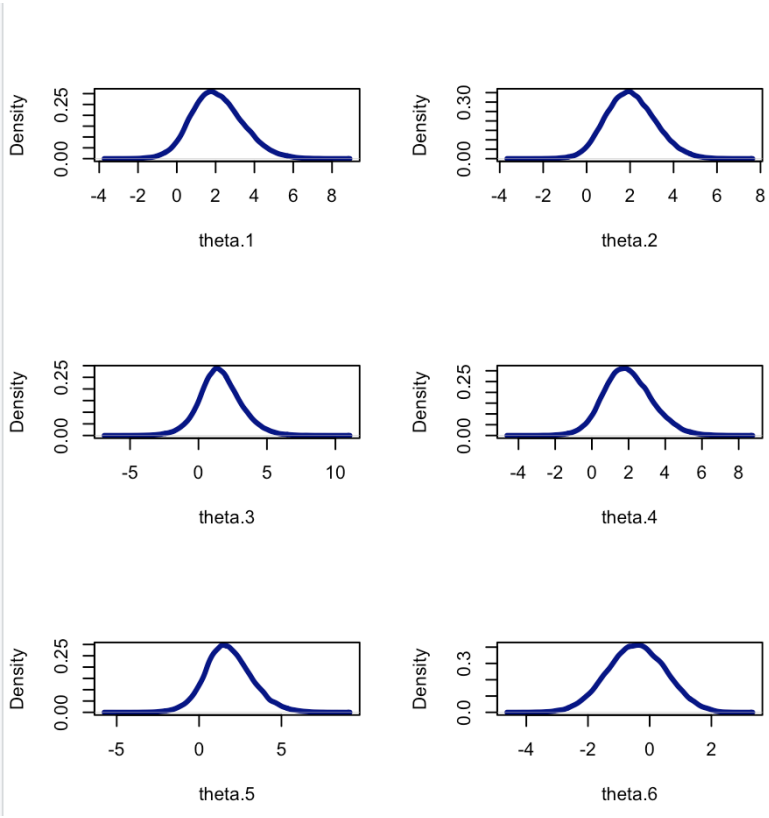
C is chosen large enough to again avoid truncation of the posterior (but not much larger than that).

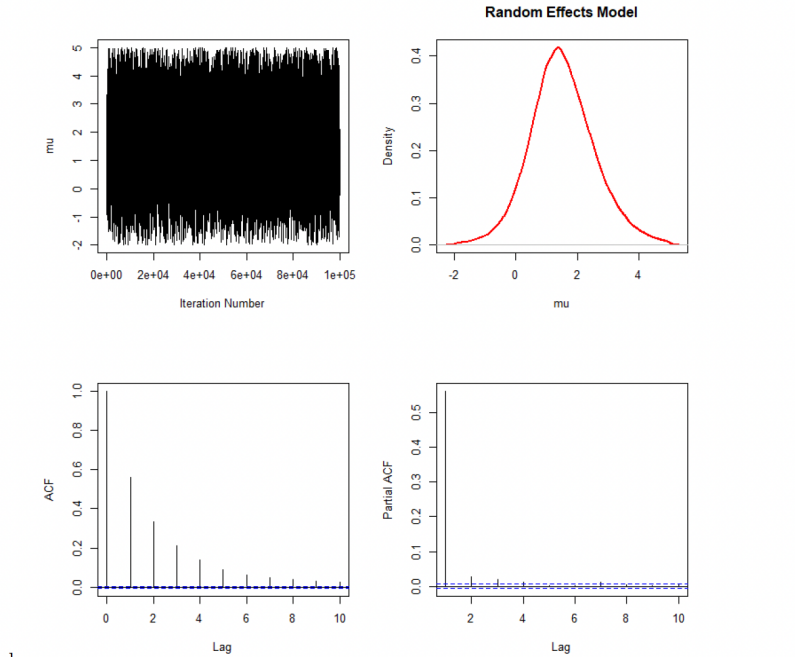
I've written `rjags` and other `R` code so that You can do the MCMC computations in this case study, and posted it on the `Pages` tab of the course `Canvas` page; the file is called

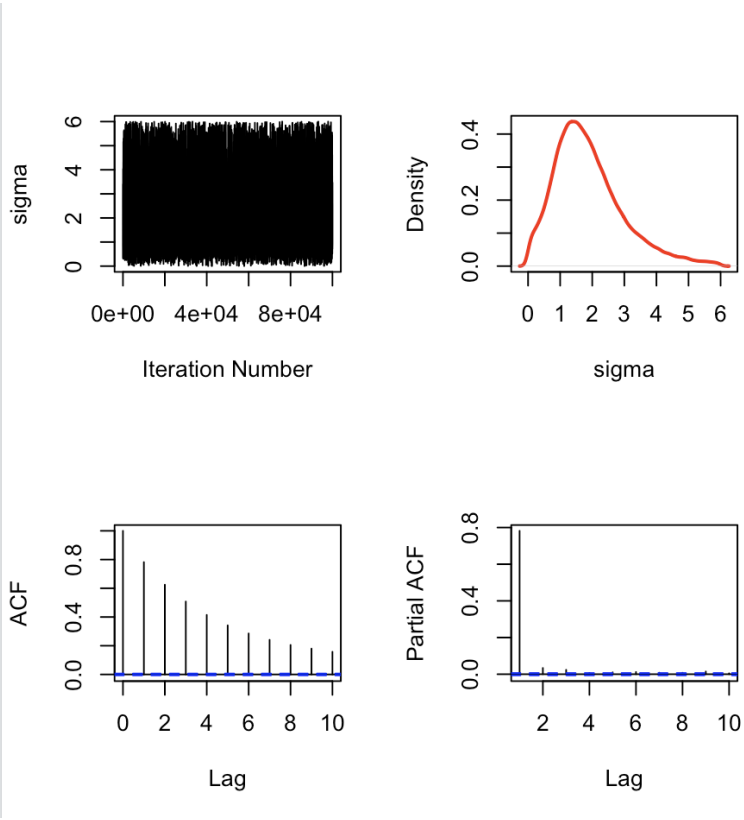
`rjags and other R code for MCMC calculations in THT 2 problem 2(B)`

Based on the likelihood visualization earlier in this problem, I chose $(A, B, C) = (-2, 5, 6)$ in the prior specification. Download the `.txt` file just mentioned, run parts (0)–(10) of my code (or an equivalent program in some other language), stopping at each place where stopping is suggested, and examine the output; make PDF files of all plots the code produces and include them in Your solutions.

(g) ***90 total points for this part of this problem*** Interpreting the MCMC output:


Density
theta.1
theta.2
theta.3
theta.4
theta.5
theta.6


Random Effects Model
mu
Iteration Number
Density
ACF
Lag
Partial ACF



(i) Use the output from Your MCMC code-running to complete Table 5 by filling in the blank entries; answering the questions below will also involve extracting additional numbers from the output. ***[10 points]***

Solution:

Refer Table 5 for the updated values.

(ii) The marginal story for μ:

(∗) Compare the posterior mean for μ with its maximum-likelihood (ML) counterpart; then compare the posterior SD for μ with the two ML standard errors, one likelihood-based and the other from empirical Bayes considerations. ***[10 points]***

(∗∗) Research on hierarchical models with random effects, such as model (18), has shown that Bayes and ML findings will either be similar (when k is large) or the ML approach will often underestimate uncertainty when it differs from Bayes. Does the second of those two possibilities appear to have happened here? Explain briefly. ***[10 points]***

Solution :

The marginal distribution for μ suggests that the posterior from MCMC and the maximum likelihood estimate (MLE) should closely align when the sample size is considerable. The likeness arises because the likelihood dominates the Bayesian posterior in large samples, making it akin to the MLE. The empirical Bayes method,

however, might yield a slightly altered μ estimate due to its variance-accounting shrinkage effect, pulling extremities towards the common mean.

(*) The posterior mean for μ and its maximum likelihood estimate are both extremely similar, while the standard error based on empirical Bayes and information based are .2 less than the standard deviation shown in Bayesian posterior.
(**) This implies that the ML approach has underestimated the uncertainty as compared to Bayes, this is likely due to the small sample size we have to work with.

(iii) The marginal story for σ:

($*$) Compare the posterior mean for σ with its ML counterpart; are they close enough that it doesn't matter which one You would report in a research article or white paper for a client? ***[10 points]***

($**$) Extract the 99.9% Bayesian posterior interval for σ from the output and report it here. ***[10 points]***

($***$) Compute the large-sample-approximate 99.9% confidence interval for σ from maximum likelihood, thereby showing that it has embarrassed itself by going negative. ***[10 points]***

($****$) Focusing on the Bayesian interval, if the Devil's Advocate (let's say female, to have a pronoun) said to You, "I think that σ is actually 0 in the population of {randomized controlled trials that could have been run in the late 1980s in Europe and the U.S. to compare aspirin with placebo for patients who have had a heart attack}, and the only reason You got something different from 0 was that the 6 studies in Your meta-analysis were unlucky," would You agree with her? Does this mean that σ is statistically significantly different from 0? Explain briefly. ***[10 points]***

Solution :

Discrepancies in σ between the MCMC output and its ML counterpart can stem from the Bayesian posterior embedding prior information, potentially altering the estimate, especially with an informative prior or a modest sample size. Furthermore, in hierarchical models, the MLE may underrate uncertainty, particularly with significant between-study heterogeneity.

(*) The posterior means for σ are not close enough that it doesn't matter which one we would report, they differ significantly, maximum likelihood gives us an estimate only $\frac{2}{3}$rds that given by the Bayesian approach.

(**) Our code gives us $(0.004579545, 5.929342194)$ as our 99.9% interval for the Bayesian posterior.

(***) The large sample approximate 99.9% confidence interval for our maximum likelihood is computed by taking our estimate 1.237 and subtracting our standard error multiplied by ~ 3.291 doing this gives us a lower bound of -0.9976.

(****) In discussion with the devil's advocate, we would state that in the Bayesian

Table 6: *DIC comparison of the fixed effects and random effects models in the aspirin meta-analysis.*

Model	Mean Deviance	Complexity Penalty	*DIC*
Fixed Effects	**27.0**	1.0	**28.1**
Random Effects	21.6	**4.056**	25.7

approach we have a 99.9% posterior interval stating that σ is non zero.

(iv) Substantive conclusions about aspirin:

(∗) Show (by extracting the relevant number from Your output) that, conditional on model (18) and the prior used to produce Your output, the posterior probability that low-dose aspirin would be beneficial, if used in the population $\mathcal{P}$ identified just above item (a) in this problem, is about 93%. ***[10 points]***

(∗∗) Is this standard of evidence strong enough for You personally to recommend the use of low-dose aspirin to prevent future heart attacks and strokes in $\mathcal{P}$? Briefly explain Your reasoning. (There is no single right answer to this question.) ***[10 points]***

Solution :

Regarding aspirin's implications, the posterior probability that low-dose aspirin is beneficial—derived from the model and prior used—implies a substantial likelihood of mortality reduction in the population identified within the meta-analysis, indicating a positive effect of aspirin. This result provides substantial evidence for clinical consideration.

print(random.effects.posterior.probability.aspirin.is.beneficial) + mean(random.effects.positive.effect.star)) [1] 0.9345

(*) From my perspective, the current body of evidence does not sufficiently support the widespread recommendation of low-dose aspirin for the secondary prevention of heart attacks. The implications drawn from these six studies suggest that further research—preferably larger and more comprehensive—is warranted. Such research is necessary to confirm the benefits of aspirin without inadvertently causing harm to individuals recovering from heart attacks. A 93% probability does not meet my threshold for advocating that aspirin be universally administered post-myocardial infarction.

(**) Given the significant mortality rates associated with heart attacks, it appears more prudent to explore alternative interventions that could be more impactful, considering that the number needed to treat to save one life with low-dose aspirin is relatively high.

(h) ***50 total points for this part of this problem*** Finally, let's make a formal comparison of the random-effects model (studied above in the rest of this problem) with the following

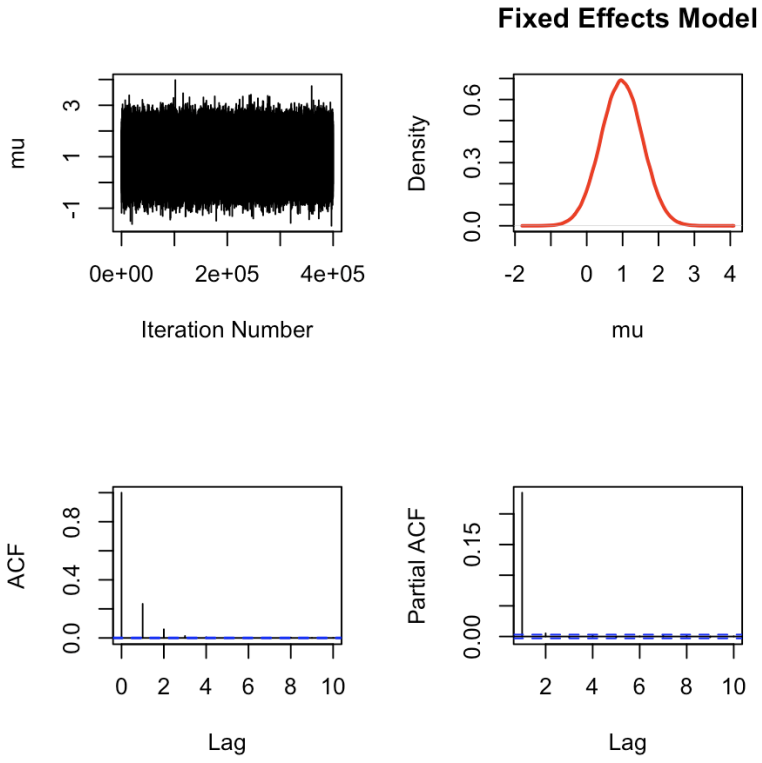
fixed-effects (FE) model for $(i = 1, \ldots, k)$:

$$\begin{aligned} (\mu \,|\, [PM]\, \mathcal{B}) &\sim p(\mu \,|\, [PM]\, \mathcal{B}) \\ (y_i \,|\, [SM\colon \mathbb{N}]\, \mu\, V_i\, \mathcal{B}) &\overset{\text{I}}{\sim} N(\mu, V_i)\,. \end{aligned} \tag{28}$$

We'll be using the Bayesian model comparison method called *DIC* (the *Deviance Information Criterion*), discussed in class as one of several such methods (and one that's suitable for working with random effects models). In `rjags` *DIC* is referred to as the *penalized deviance*, the measure of model complexity that *DIC* uses is called the *penalty* term, the measure of model lack-of-fit is referred to as the *mean deviance*, and *DIC* works by combining the two terms to resolve the tradeoff between complexity and lack of fit.

As was true in the discussion above just before part (g), from context we also want a $[LI]$ prior for μ in the FE model; in the code for the FE model I've used the same $[LI]$ choice as in the random-effects model.

(i) By examining the random effects model equations (18), briefly explain why the fixed effects model in (28) is a special case of (18) in which it's assumed that $\sigma = 0$. ***[10 points]***



In the scenario where σ is null, equation 15 simplifies to equation 25, since the variance component σ that accounts for between-study heterogeneity is no longer present. Consequently, the model reduces to a single-layer normal distribution centered around μ, effectively collapsing the hierarchical structure. In this model, μ emerges as the

sole stochastic element, a condition that arises exclusively when there is an absence of between-study variance.

(ii) Interpreting the *DIC* results:

(∗) Run the final block of code (section (11) in the `rjags` code file) to get *DIC* values for the fixed effects and random effects models, and use your output to fill in the missing (blank) entries in Table 6. ***[10 points]***

Refer the **filled** table 6.

(∗∗) Which model is more complex? Which model fits better? Explain briefly. ***[10 points]***
COMPLEX: Fixed Effects BEST-FIT: Random Effects

(∗∗∗) Bearing in mind that *DIC* is set up so that smaller values indicate better models, which of the two models is more strongly supported by the *DIC* evidence here? ***[10 points]***

The graphical representations we've generated lend support to the random effects model. Additionally, we observe considerably lower values of mean deviance and penalized deviance relative to the fixed effects model. It is therefore reasonable to assert that between-study heterogeneity exerts a significant influence on the outcomes of our statistical analyses.

(∗∗∗∗) Does this agree with your conclusions about between-study heterogeneity in the earlier parts of this problem? Explain briefly. ***[10 points]***

The findings corroborate our prior determinations; the probability of σ being null is remote. Omitting its influence from consideration would be methodologically remiss.