Prof. David Draper
University of California, Santa Cruz
Department of Statistics
Baskin School of Engineering
Winter 2022

STAT 206 (Applied Bayesian Statistics)

Take-Home Test 2

(please watch email and `Canvas` for the final due date)

Name: Kevin Guillen

Here are the (process) ground rules: this test is open-book and open-notes, and consists of two problems (true/false and calculation); ; **each of the 6 true/false questions is worth**[1] ***10 points*****, and the calculation problem is worth** ***310 total points*****, for a total of** ***370 points***.

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

On each problem, the graders and I mentally start everybody out at −0 (i.e., with a perfect score), and then you accumulate negative points for incorrect answers and/or reasoning, or parts of problems left blank.

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TA (Jacob Fontana). The intent is that the course lecture notes and readings should be sufficient to provide you with all the guidance you need to solve the problems posed below, but you may use other written materials (e.g., the web, journal articles, and books other than those already mentioned in the readings), **provided that you cite your sources thoroughly and accurately**; you will lose (substantial) credit for, e.g., lifting blocks of text directly from `wikipedia` and inserting them into your solutions without full attribution.

If it's clear that (for example) two people have worked together on a part of a problem that's worth 20 points, and each answer would have earned 16 points if it had not arisen from a collaboration, then each person will receive 8 of the 16 points collectively earned (for a total score of 8 out of 20), and I reserve the right to impose additional penalties at my discretion. If you solve a problem on your own and then share your solution with anyone else, you're just as guilty of illegal collaboration as the person

[1] Throughout the test, I've tried to be completely clear about the location of each sub-part of each problem by surrounding the possible points with boxes and putting the text inside the boxes in bold italic font.

who took your solution from you, and both of you will receive the same penalty. This sort of thing is necessary on behalf of the many people who do not cheat, to ensure that their scores are meaningfully earned. In the AMS graduate class in 2012 mentioned above, five people failed the class because of illegal collaboration; don't let that happen to you.

In class I've demonstrated numerical work in `R`; you can (of course) make the calculations and plots requested in the problems below in any environment you prefer (e.g., `Matlab`, ...). To avoid plagiarism, if you end up using any of the code I post on the course web page or generate during office hours, at the beginning of your Appendix (see below) you can say something like the following:

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

Those of You who are using `LaTeX` or some other word-processing environment to prepare Your solutions can stick quote blocks below each question, into which You can type Your answers (I suggest that You use **bold** or *italic* font to distinguish Your solutions from the questions). If You're submitting Your answers in longhand, which is perfectly acceptable, You can just write them out on separate sheets of paper, making sure that the grader can easily figure out which chunk of text is the solution to which part of which problem.

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

> ***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

[***60 total points:*** $\boxed{\textbf{\textit{10 points each}}}$] 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 sampling model $(Y_i \,|\, \boldsymbol{\theta}\, \mathcal{B}) \stackrel{\text{IID}}{\sim} p(y_i \,|\, \boldsymbol{\theta}\, \mathcal{B})$ 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 $1 \leq k < \infty$ and $\mathcal{B}$ summarizes Your background information; a Bayesian analysis with the same sampling model would add a prior distribution layer of the form $(\boldsymbol{\theta} \,|\, \mathcal{B}) \sim p(\boldsymbol{\theta} \,|\, \mathcal{B})$ to the hierarchy. 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 large and (b) $p(\boldsymbol{\theta})$ is a low-information (LI) prior, but the theorem does not provide guidance on how large n needs to be for its conclusion to hold in any specific sampling model. $\boxed{\textbf{\textit{10 points}}}$

Solution. This is **True**, it follows from what we covered today in lecture 2/23. We see it in practice when we compare the likelihood PDF with the posterior PDF when we have low information priors. □

(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 the document camera notes from 20 Jan 2022), 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 $\bar{Y}$, 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. This is **False**.
Well first we can send N to infinity by duplicating the population data set and adding it to itself over and over, this will lead to no effect on the long run distribution of $\bar{Y}_n$ in the repeated sampling data set.
Now consider if $n = 5$ a relatively small sample size. This means that for each sample the mean of that sample can only be,

$$0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0$$

implying then that if we take more repeated samples (increasing M), the long run distribution of these means will never get any closer to the normal curve no matter how large M gets as (N, n) remain fixed. It will just get closer to the true PMF for $\bar{Y}_n$ given the population proportions of 1's and 0's. With small n this will be a spike plot with $(n + 1)$ spikes.
By the Central Limit Theorem, as long as the population standard deviation remains positive and finite, as n increase the PMF or PDF for $\bar{Y}_n$ will approach normality.
In order to actually see what is going on though, we need n and M to increase in the diagram for the CLT to emerge. This is what makes this statement **True**.
□

(C) The ability to express Your sampling distribution 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. Based on what we learned in class on 2/22 this is **True**. □

(D) When the sampling model is a regular[2] parametric family $p(\boldsymbol{y} \,|\, \boldsymbol{\theta}\, \mathcal{B})$, where $\boldsymbol{\theta}$ is a vector of length $1 < k < \infty$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, for large n 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 $O\big(\frac{1}{n^2}\big)$. ***10 points***

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

Solution. This is **False** based on what we learned in class on 3/03. To make it true we simply have to change the bias of $\hat{\theta}_{MLE}$ to $O\left(\frac{1}{n}\right)$, NOT $O\left(\frac{1}{n^2}\right)$ □

(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. Very clearly **False**. One way to make this statement true is simply replacing the two instances of "easier" with "harder". □

(F) When Your sampling model has n observations and a single parameter θ (so that $k = 1$), if the sampling model is regular[2], 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 variance $\hat{V}_{RS}\left(\hat{\theta}_{MLE}\right)$ is $O\left(\frac{1}{n}\right)$.

10 points

Solution. This is **True** from what we saw in lecture on 3/02, what we saw happens in CS2 actually happen in general. □

2 Calculation (A)

[***105 total points***] From 29–31 Oct 2020, a sample survey was conducted by the highly-regarded polling firm *SurveyUSA*[3] 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.

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

[3] On 2 Nov 2021 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.

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[4]. In this situation people say that $\boldsymbol{N}$ follows the Multinomial distribution with parameters n and $\boldsymbol{\theta}$, which is defined as follows: $(\boldsymbol{N} \,|\, n\, \boldsymbol{\theta}\, \mathcal{B}) \sim \text{Multinomial}(n, \boldsymbol{\theta})$ iff

$$P(N_1 = n_1, \ldots, N_k = n_k \,|\, n\, \boldsymbol{\theta}\, \mathcal{B}) = \left\{ \begin{array}{cc} \frac{n!}{n_1!\, n_2! \cdots n_k!}\, \theta_1^{n_1}\, \theta_2^{n_2} \cdots \theta_k^{n_k} & \text{if } n_1 + \cdots + n_k = n \\ 0 & \text{otherwise} \end{array} \right\}, \qquad (1)$$

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).

[4]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).

(a) [**15 total points** *for this sub-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. If we try to visualize the raw data set it would look something along the lines of,

$$1,264 = n \begin{cases} \begin{bmatrix} T \end{bmatrix} \\ \begin{bmatrix} B \end{bmatrix} \\ \begin{bmatrix} T \end{bmatrix} \\ \begin{bmatrix} O \end{bmatrix} \\ \begin{bmatrix} \vdots \end{bmatrix} \end{cases}$$

We see it is **qualitative** since who people are voting for isn't reduced to numbers (voting for Biden, Trump, etc), which is why this data isn't **quantitative**. Similarly this data is **categorical** since it is synonymous to qualitative. It is **nominal** since there is no ordering in the categories, which is why it isn't **ordered categorical**, it isn't **dichotomous** since there is more than 2 category label. The rest do not apply here since those terms don't make sense when trying to talk about qualitative data, they only work when talking about quantitative data which isn't present here.
The numbers $\mathbf{n} = (n_1, n_2, n_3)$ are not raw data values since we are taking the raw data above, and organizing them into their categories and using number of votes to represent each category (how many people of that sample will be voting for who) which is why it is instead a summary of the raw data vector. □

(b) [**5 total points** *for this sub-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: the Binomial distribution 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

$$(X \,|\, n\,\theta\,\mathcal{B}) \sim \text{Binomial}(n,\theta) \text{ iff } P(X = x \,|\, n\,\theta\,\mathcal{B}) = \left\{ \begin{array}{cc} \binom{n}{x} \theta^x (1-\theta)^{n-x} & \text{for } x = 0,\ldots,n \\ 0 & \text{otherwise} \end{array} \right\} . \quad (2)$$

Briefly and carefully explain why the correspondence between equation (2) and {a version of equation (1) with $k = 2$} is as in Table 1. **5 points**

Solution. To show how the Multinomial is indeed a generalization of the the Binomial we will consider the table below and how if we select our notations correctly the Binomial will look Multinomial just $k = 2$,

Category	Binomial Count	Multinomial Count
Success	x	n_1
Failure	$n - x$	n_2
Total	n	n

Now if we write equation (1) in the case that $k = 2$,

$$P(N_1 = n_1, N_2 = n_k \,|\, n\theta\mathcal{B}) = \left\{ \begin{array}{cc} \frac{n!}{n_1! n_2!} \theta_1^{n_1} \theta_2^{n_2} & \text{if } n_1 + n_2 = n \\ 0 & else \end{array} \right\}$$

and the Binomial equation was the exactly the same except for the calculation which was,

$$\binom{n}{x} \theta^x (1-\theta)^{n-x}$$

where if we algebraiclly expand we get,

$$\frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x}$$

So applying what we have from our table we see $x = n_1$ and $(n - x) = n_2$, and we must have $\theta = \theta_1$ and $(1 - \theta) = \theta_2$. Meaning equation (2) is just a specialization of equation (1). □

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 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

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

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***, 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*:

$$(Y_i \,|\, \boldsymbol{C}\,\mathcal{B}) \overset{\text{IID}}{\sim} \text{Categorical}(\boldsymbol{C}) \quad \rightarrow \quad p(y_i \,|\, \mathcal{B}) = \left\{ \begin{array}{cc} \theta_1 & \text{if} \quad y_i = \text{'Biden'} \\ \theta_2 & \quad\;\; y_i = \text{'Trump'} \\ \theta_3 & \quad\;\; y_i = \text{'Other'} \\ 0 & \quad\;\; \text{otherwise} \end{array} \right\} , \tag{3}$$

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 sampling model 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.

(c) [***20 total points*** *for this sub-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

$$\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\,\mathcal{B}) = c_+ \prod_{j=1}^{k} \theta_j^{n_j} \tag{4}$$

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

$$\ell\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\,\mathcal{B}) = c + \sum_{j=1}^{k} n_j \, \log \theta_j \,, \tag{5}$$

where c is an arbitrary real constant. ***[5 points]***

> ***Solution***. Let's consider the case where $k = 1$ our likelihood function was simply,
>
> $$\ell(\theta \,|\, \boldsymbol{y}\mathcal{B}) = C_+ P(\boldsymbol{y} \,|\, \theta\mathcal{B})$$

and we have no issues arise when we treat θ as a $k-$dimensional vector for $k > 1$ to get,

$$\ell(\boldsymbol{\theta} \,|\, \boldsymbol{y}\mathcal{B}) = C_+ P(\boldsymbol{y} \,|\, \boldsymbol{\theta}\mathcal{B}).$$

So applying this to our problem we get,

$$\begin{aligned}\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\mathcal{B}) &= c_+ P(\boldsymbol{n} \,|\, \boldsymbol{\theta}\mathcal{B}) \\ &= c_+ \frac{n!}{n_1! n_2! \dots n_k!} \theta_1^{n_1} \theta_2^{n_2} \dots \theta_k^{n_k}\end{aligned}$$

Now because the likelihood function is a function of theta for a fixed $\boldsymbol{\theta}$ for a fixed $\boldsymbol{n}$ we can let the constant c_+ absorb the fraction in the equation above since that fraction has noting to do with $\boldsymbol{\theta}$, this gives us,

$$\begin{aligned}\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\mathcal{B}) &= c_+ \frac{n!}{n_1! n_2! \dots n_k!} \theta_1^{n_1} \theta_2^{n_2} \dots \theta_k^{n_k} \\ &= c_+ \prod_{j=1}^{k} \theta_j^{n_j}\end{aligned}$$

as desired.
Now to obtain the log-likelihood function we simply do the following,

$$\begin{aligned}\ell\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\mathcal{B}) &= \log(\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\mathcal{B})) \\ &= \log(c_+ \prod_{j=1}^{k} \theta_j^{n_j}) \\ &= \log(c_+) + \log(\prod_{j=1}^{k} \theta_j^{n_j}) \\ &= c + \sum_{j=1}^{k} \log(\theta_j^{n_j}) \\ &= c + \sum_{j=1}^{k} n_j \log(\theta_j)\end{aligned}$$

which matches the equation given in the problem statement. □

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 $\ell\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\,\mathcal{B})$ 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* (Jacob and I will show you how to do this in office hours if You forget or don't know, because `Wolfram Alpha` is useless here) ***[5 points]***,

Solution. To break this down we are trying to maximize,

$$\ell\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\mathcal{B}) = c + \sum_{j=1}^{k} n_j \log(\theta_j)$$

under the constraint that $\sum_{j=1}^{k} \theta_j = 1$. To use Lagrange multipliers though, our constraint will be $g(\boldsymbol{\theta}) = \sum_{j=1}^{k} -1 = 0$ in order to form our Lagrangian function,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \lambda) &= \ell\ell_C(\boldsymbol{\theta} \,|\, \boldsymbol{n}\mathcal{B}) - \lambda g(\boldsymbol{\theta}) \\ &= c + \sum_{j=1}^{k} n_j \log(\theta_j) - \lambda(\sum_{j=1}^{k} \theta_j - 1)\end{aligned}$$

Now taking the a partial derivative with respect to each θ_i and to λ, then setting it to 0 we get,

$$\begin{aligned}\frac{d}{d\theta_1}\mathcal{L}(\boldsymbol{\theta}, \lambda) &= \frac{n_1}{\theta_1} - \lambda = 0 \\ &\vdots \\ \frac{d}{d\theta_k}\mathcal{L}(\boldsymbol{\theta}, \lambda) &= \frac{n_k}{\theta_k} - \lambda = 0 \\ \frac{d}{d\lambda}\mathcal{L}(\boldsymbol{\theta}, \lambda) &= -(\sum_{j=1}^{k} \theta_j - 1) = 0\end{aligned}$$

Solving the first k equations for n_i we get,

$$\begin{aligned}n_1 &= \theta_1\lambda \\ n_2 &= \theta_2\lambda \\ &\vdots \\ n_k &= \theta_k\lambda\end{aligned}$$

we know the last equation holds true if and only if $\theta_1 + \theta_2 + \cdots + \theta_k = 1$. Recall that $n = n_1 + n_2 + \cdots + n_k$. Putting this together the above we have then,

$$\begin{aligned}n &= n_1 + n_2 + \cdots + n_k \\ n &= \theta_1\lambda + \theta_2\lambda + \cdots + \theta_k\lambda \\ n &= \lambda(\theta_1 + \theta_2 + \cdots + \theta_k) && \text{Recall our constraint} \\ n &= \lambda\end{aligned}$$

Therefore we have then that,

$$\theta_i = (\hat{\theta}_i)_{MLE} = \frac{n_i}{\lambda} = \frac{n_i}{n}$$

□

(iii) Build the constraint directly into the likelihood function: since $\sum_{j=1}^{k} \theta_j = 1$, we can

write $\theta_k = 1 - \sum_{j=1}^{k-1} \theta_j$ and define

$$\ell_U(\theta_1, \ldots, \theta_{k-1} \,|\, \boldsymbol{n}\,\mathcal{B}) = c_+ \left(\prod_{j=1}^{k-1} \theta_j^{n_j} \right) \left(1 - \sum_{j=1}^{k-1} \theta_j \right)^{n_k} \tag{6}$$

(here the subscript U stands for *Unconstrained*), from which

$$\ell\ell_U(\theta_1, \ldots, \theta_{k-1} \,|\, \boldsymbol{n}\,\mathcal{B}) = \sum_{j=1}^{k-1} n_j \, \log \theta_j + n_k \log \left(1 - \sum_{j=1}^{k-1} \theta_j \right) . \tag{7}$$

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

$$\frac{\partial}{\partial\, \theta_j} \, \ell\ell_U(\theta_1, \ldots, \theta_{k-1} \,|\, \boldsymbol{n}\,\mathcal{B}) = \frac{n_j}{\theta_j} - \frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i} \tag{8}$$

[5 points]

Solution. Well first when we take the derivative of a constant to a variable it goes away. Next if we expand the first summation we get,

$$\sum_{j=1}^{k-1} n_j \log \theta_j = n_1 \log(\theta_1) + \cdots + n_{k-1} \log(\theta_{k-1})$$

so taking the partial derivative of this summation with respect to θ_j will yield,

$$n_j \cdot \frac{1}{\theta_j}.$$

Now taking the partial derivative with respect to θ_j of $n_k \log(1 - \sum_{i=1}^{k-1} \theta_i)$ we get,

$$\frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i} \cdot -1$$

which just follows from the rules of differentiating the log function and chain rule. Putting all this together we get,

$$\frac{\partial}{\partial\, \theta_j} \, \ell\ell_U(\theta_1, \ldots, \theta_{k-1} \,|\, \boldsymbol{n}\,\mathcal{B}) = \frac{n_j}{\theta_j} - \frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i}$$

□

The MLE for $(\theta_1, \ldots, \theta_{k-1})$ may now be found by setting $\frac{\partial}{\partial\, \theta_j} \, \ell\ell_U(\theta_1, \ldots, \theta_{k-1} \,|\, \boldsymbol{n}\,\mathcal{B}) = 0$ for $j = 1, \ldots, (k-1)$ 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

$$\left\{ \frac{n_1}{\theta_1} - \frac{n_3}{1 - \theta_1 - \theta_2} = 0, \quad \frac{n_2}{\theta_2} - \frac{n_3}{1 - \theta_1 - \theta_2} = 0 \right\} \tag{9}$$

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

$$\hat{\boldsymbol{\theta}}_{MLE} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}\right) . \tag{10}$$

[5 points]

Solution. First we can cross multiply the left hand side of each equation given to obtain,

$$\{n_1 - n_1\theta_1 - n_1\theta_2 - n_3\theta_1 = 0,\ n_2 - n_2\theta_1 - n_2\theta_2 - n_3\theta_2 = 0\} .$$

Now we can solve the first equation for n_1 and the second one for n_2 to get,

$$\{n_1 = (n_1 + n_3)\theta_1 + n_1\theta_2,\ n_2 = n_2\theta_1 + (n_2 + n_3)\theta_2\}$$

We know that $n_1 + n_2 + n_3 = n$ so we can write n_3 as $n_3 = n - n_2 - n_1$ and plug it in to obtain,

$$\{n_1 = (n - n_2)\theta_1 + n_1\theta_2,\ n_2 = n_2\theta_1 + (n - n_1)\theta_2\}$$

which gives us a system of equations $Ax = y$ where,

$$A = \begin{bmatrix} n - n_2 & n_2 \\ n_1 & n - n_1 \end{bmatrix},\ x = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix},\ y = \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}$$

meaning we get the solution for x through $x = A^{-1}y$.
So first we calculate $A^{-1} = \frac{1}{|A|}\begin{bmatrix} n - n_1 & -n_2 \\ -n_1 & n - n_2 \end{bmatrix}$ now let's proceed,

$$|A| = (n - n_2)(n - n_1) - n_2 n_1 = n^2 - n_2 n - n n_1$$

so we have,

$$\begin{aligned} x &= \frac{1}{n^2 - n_2 n - n n_1}\begin{bmatrix} n - n_1 & -n_2 \\ -n_1 & n - n_2 \end{bmatrix}\begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \\ &= \frac{1}{n^2 - n_2 n - n n_1}\begin{bmatrix} n n_1 - n_1^2 - n_1 n_2 \\ n n_2 - n_2^2 - n_2 n_1 \end{bmatrix} \\ &= \frac{1}{n(n - n_2 - n_1)}\begin{bmatrix} n_1(n - n_2 - n_1) \\ n_2(n - n_2 - n_1) \end{bmatrix} \\ &= \begin{bmatrix} \frac{n_1}{n} \\ \frac{n_2}{n} \end{bmatrix} \end{aligned}$$

So we have $\theta_1 = \frac{n_1}{n}$ and $\theta_2 = \frac{n_2}{n}$. Now to solve for θ_3 we recall out constraint that,

$$\begin{aligned} \theta_1 + \theta_2 + \theta_3 &= 1 \\ \theta_3 &= 1 - \theta_2 - \theta_1 \\ \theta_3 &= 1 - \frac{n_2}{n} - \frac{n_1}{n} \\ \theta_3 &= \frac{n - n_2 - n_1}{n} = \frac{n_3}{n} \end{aligned}$$

So we have,

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

□

(The result for general k, of course, is[5] 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, $\hat{\gamma}_{MLE} = (\hat{\theta}_1 - \hat{\theta}_2)$.)

(d) [**10 total points** *for this sub-problem*] 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

$$\hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n} & -\frac{\hat{\theta}_1\,\hat{\theta}_2}{n} & -\frac{\hat{\theta}_1\,\hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1\,\hat{\theta}_2}{n} & \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n} & -\frac{\hat{\theta}_2\,\hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1\,\hat{\theta}_3}{n} & -\frac{\hat{\theta}_2\,\hat{\theta}_3}{n} & \frac{\hat{\theta}_3(1-\hat{\theta}_3)}{n} \end{pmatrix} . \tag{11}$$

(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 random vector with covariance matrix $\hat{\boldsymbol{\Sigma}}$ and $\gamma = \boldsymbol{a}^T\boldsymbol{\theta}$ for some vector $\boldsymbol{a}$ of constants, then in repeated sampling

$$\hat{V}(\hat{\gamma}) = \hat{V}\left(\boldsymbol{a}^T\hat{\boldsymbol{\theta}}\right) = \boldsymbol{a}^T\hat{\boldsymbol{\Sigma}}\,\boldsymbol{a}\,. \tag{12}$$

[5 points]

Solution. We'll we have that for any j, $\hat{V}(\hat{\theta}_j) = \dfrac{\hat{\theta}_j(1-\hat{\theta}_j)}{n}$. Now to get the standard error for each of these $\hat{\theta}_j$ we simply square them. In other words we simply calculate,

$$\widehat{SE}(\hat{\theta}_j) = \sqrt{\hat{V}(\hat{\theta}_j)} = \sqrt{\frac{\hat{\theta}_j(1-\hat{\theta}_j)}{n}}$$

Plugging this into R we obtain

$\hat{\theta}_j$	Standard Error
$\hat{\theta}_1$	0.5209486
$\hat{\theta}_2$	0.4379447
$\hat{\theta}_3$	0.04110672

Since we have $\hat{\gamma} = (\hat{\theta}_1 - \hat{\theta}_2)$ to calculate the standard error of it would be to calculate the standard error of $\hat{\theta}_1 - \hat{\theta}_2$ which is the square root of variance. Mathematically we have,

$$\widehat{SE}(\hat{\gamma}) = \widehat{SE}(\hat{\theta}_1 - \hat{\theta}_2) = \sqrt{\hat{V}(\hat{\theta}_1 - \hat{\theta}_2)}.$$

[5]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.

To do this let us first calculate the variance between the difference of two random variables. This though we know from 131 to be,

$$\hat{V}(\hat{\theta_1} - \hat{\theta_2}) = \hat{V}(\hat{\theta_1}) + \hat{V}(\hat{\theta_2}) - 2\hat{C}(\hat{\theta_1}, \hat{\theta_2}).$$

These variances and covariance are given by the estimated covariance matrix above specifically by (1,1) and (1,2) cells. So plugging this in we get,

$$\begin{aligned}\hat{V}(\hat{\theta_1} - \hat{\theta_2}) &= \hat{V}(\hat{\theta_1}) + \hat{V}(\hat{\theta_2}) - 2\hat{C}(\hat{\theta_1}, \hat{\theta_2}) \\ &= \frac{\hat{\theta_1}(1-\hat{\theta_1})}{n} + \frac{\hat{\theta_2}(1-\theta_2)}{n} - 2\left(-\frac{\hat{\theta_1}\hat{\theta_2}}{n}\right)\end{aligned}$$

Now plugging this all into R it all evaluates to,

$$\hat{V}(\hat{\theta_1} - \hat{\theta_2}) = 0.000753$$

So now we just take the square root of this obtain the standard error, which works out to be,

$$\widehat{SE}(\hat{\gamma}) = \sqrt{\hat{V}(\hat{\gamma})} = \sqrt{0.000753} = 0.027433$$

□

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

$$\hat{\gamma}_{MLE} - \Phi^{-1}(1-\alpha) \cdot \widehat{SE}\left(\hat{\gamma}_{MLE}\right) \tag{13}$$

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 (13).

(ii) Finally, use Your estimated SE for $\hat{\gamma}$ to construct an approximate (large-sample) 99.9% LCB for γ *[5 points]*. Was Biden ahead of Trump at the point when the survey was conducted by an amount that was large in *practical* terms? Was Biden's lead at that point *statistically* significant at the 99.9% level? Explain briefly. ***[5 points]***

Solution. We know $\gamma_{\hat{MLE}} = (\hat{\theta_1} - \hat{\theta_2}) = 0.0830$ and we worked out our standard error of it to be 0.0274. So using R and plugging in what we have into the given equation we get our 99.9% LCB for γ to be,

$$0.0830 - (3.08) \cdot (0.0274) = -0.0018$$

Meaning this lead that Biden had over Trump is not quite statistically significant (since it has a negative lower bound), but it was in practical terms, since γ worked out to be around 8 percentage points and that much of a lead in a presidential debate is significant. The lead did become statistically significant though at the 99.8% confidence interval level. □

(e) **[10 total points]** *for this sub-problem]* Looking back at equation (4), if a conjugate prior exists for the Multinomial likelihood it would have to be of the form

> 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 *Dirichlet*($\boldsymbol{\alpha}$) 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:

$$p(\boldsymbol{\theta} \,|\, \mathbb{D}) = c \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \,; \tag{14}$$

here $\mathbb{D}$ stands for the Dirichlet prior distribution assumption, which is not part of $\mathcal{B}$.

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

$$\left\{ \begin{array}{rcl} (\boldsymbol{\theta} \,|\, \mathbb{D}\,\mathcal{B}) & \sim & \text{Dirichlet}(\boldsymbol{\alpha}) \\ (\boldsymbol{N} \,|\, \boldsymbol{\theta}\, n\, \mathcal{B}) & \sim & \text{Multinomial}(n, \boldsymbol{\theta}) \end{array} \right\} \longrightarrow (\boldsymbol{\theta} \,|\, \boldsymbol{N}\, \mathbb{D}\, \mathcal{B}) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{N}) \,. \tag{15}$$

[5 points]

> ***Solution***. Well I think this would best be explained through some algebra. We see if we take the Dirichlet prior distribution and our likelihood and their product we get,
>
> $$\begin{aligned} P(\boldsymbol{\theta} \,|\, \boldsymbol{N}\mathbb{D}\boldsymbol{\alpha}\mathcal{B}) &= \left(C_+ \theta_1^{\alpha_1 - 1} \cdot \cdots \cdot \theta_k^{\alpha_k - 1}\right) \left(C_+ \theta_1^{n_1} \cdot \cdots \cdot \theta_k^{n_k}\right) \\ &= C_+ \theta_1^{\alpha_1 + n_1 - 1} \cdot \cdots \cdot \theta_k^{\alpha_k + n_k - 1} \\ &= \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_k + n_k) \\ &= \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{N}) \end{aligned}$$
>
> which explains the problem statements, since we see that constants just combine, and the exponents get added. In other words the product of two Dirichlet distributions is another one, as we have seen before. □

(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, this would correspond to choosing all of the α_j to be positive but close to 0. **[5 points]**

> ***Solution***. This follows simply from the fact that our posterior "votes" are a sum of our prior votes (α_i) and data votes (n_i) for $i = 1, 2, \ldots, k$. So if context were to suggest a low information prior, that would be the same as our α_i's being positive, but specifically close to zero. □

(f) [**45 total points** *for this sub-problem*] Computation with the Dirichlet posterior distribution:

(i) Briefly explain why, if You have a valid way of sampling from the Dirichlet distribution, it's not necessary in this problem in fitting model (15) to do MCMC sampling: IID Monte Carlo sampling is sufficient **[5 points]**.

Solution. This is because, as we've seen in class, if we can do Monte Carlo sampling simulations fast enough, we don't need to do Markov Chain Monte Carlo sampling. □

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:

$$g_j \overset{\text{I}}{\sim} \Gamma(\alpha_j, \beta) \quad \text{and} \quad \theta_j = \frac{g_j}{\sum_{i=1}^{k} g_i}\,, \tag{16}$$

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

I've written an `R` function called `rdirichlet`, posted in the `Pages` tab in the course `Canvas` page, that implements this algorithm; the relevant file is called

`R code for making IID draws from the Dirichlet( alpha ) distribution (THT 2 problem 2(A))`

(ii) Download this `.txt` file and use my function (or an equivalent in Your favorite non-`R` environment) to generate M IID draws from the posterior distribution specified by model (15), using the *SurveyUSA* polling data and a diffuse 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 $\boldsymbol{\theta}$, also monitor $\gamma = (\theta_1 - \theta_2)$. Choose a value of M 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 0.00005, and justify Your choice. ***[15 points]***

Solution. Using the function provided to us written in R, plugging in the data with desired values, and choosing $M = 100,000$ we obtain the following results,

	θ_1	θ_2	θ_3	γ
Mean	0.52096353	0.43793924	0.04109723	0.08302429
Standard Error	0.00004431	0.00004399	0.00001766	0.00008652
Target	0.00005	0.00005	0.00005	0.00005

We see with this many draws $\boldsymbol{\theta}$ achieves the desired standard error range, but our γ does not. So to determine the needed amount of draws, let us do some algebra. Just like we implemented in the code, to calculate the Standard Error,

$$\widehat{MCSE}(\gamma) = \frac{\sigma_\gamma}{\sqrt{M}}.$$

So we can actually solve for the needed M by simply setting the $\widehat{MCSE}$ (Estimate of Monte Carlo Standard Error) to 0.00005 and solving for M,

$$0.00005 = \frac{\sigma_\gamma}{\sqrt{M}}$$
$$M = (\frac{\sigma_\gamma}{0.00005})^2$$

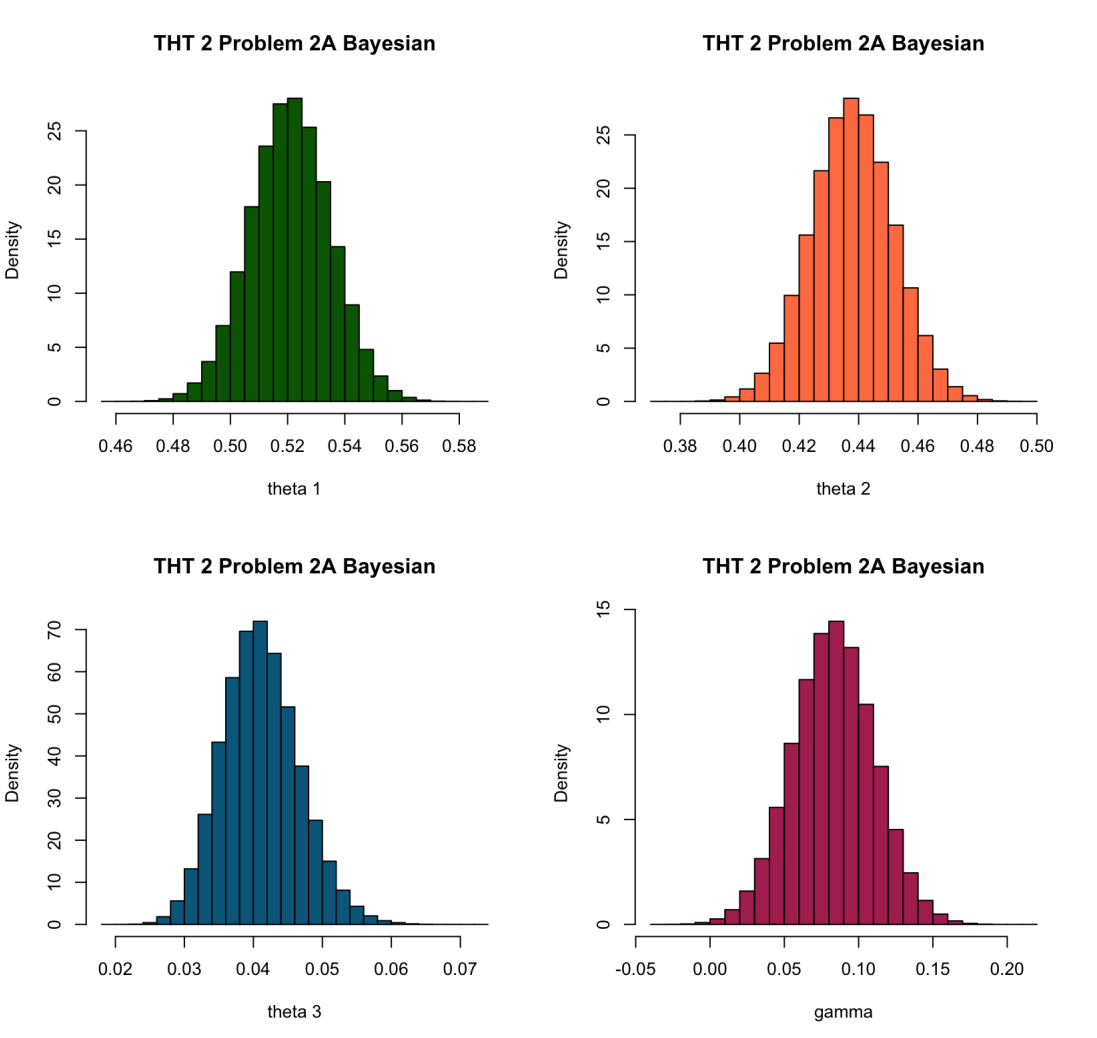
when plugging into R and using the ceiling function it works out to be, $301,630$. Now doing all these calculations again, we get,

	θ_1	θ_2	θ_3	γ
Mean	0.52100233	0.43788408	0.04111358	0.08311825
Standard Error	0.00002559	0.00002541	0.00001019	0.00004985
Target	0.00005	0.00005	0.00005	0.00005

as we see all of our θ's and γ Standard Error is within the desired range. So for this data set's case $M = 301,630$ was sufficient (Note: I didn't use the same seed that professor Draper used so my M is a little different). □

(iii) Make graphical and numerical summaries of the posterior distributions for γ and for each of the components of $\boldsymbol{\theta}$ ***[10 points]***

Solution.



□

(iv) How do Your Bayesian answers compare with those from maximum likelihood in this

problem? Explain briefly. ***[5 points]***

Solution.

Method	θ_1	θ_1	θ_2	θ_2	θ_3	θ_3	γ	γ
	Mean	SE	Mean	SE	Mean	SE	Mean	SE
(1)	0.521	0.0140	0.438	0.0139	0.0411	0.00558	0.0830	0.0274
(2)	0.521	0.0141	0.438	0.0140	0.0411	0.00557	0.0830	0.0274

Where (1) and (2) are "Likelihood" and "Bayes w/ LI prior" respectively. We see that there is an agreement between the two. These were obtained from the same code to answer the previous parts of the question which can be seen at the end of the document. □

(v) Compute a Monte Carlo estimate of $p(\gamma > 0 \,|\, \boldsymbol{N}\, \mathbb{D}\, \mathcal{B})$, which quantifies the current information about whether Biden was leading Trump in the population of all adult Americans eligible to vote, and attach a Monte Carlo standard error to Your estimate; this is the Bayesian analogue of the frequentist 99.9% LCB for γ in part (d)(ii). On the basis of this Bayesian calculation, was Biden's lead statistically significant at the 99.9% level? ***[5 points]***

Solution. The Monte Carlo estimate of the posterior probability that γ is positive is,

$$p(\gamma > 0 \,|\, \boldsymbol{N}, \mathbb{D}, \mathcal{B}) = 0.99875$$

which has a Monte Carl standard error of 0.0000644, which we calculate in a similar fashion to θ_i which is seen in the code at the end of the document.
Finally, this is indeed statistically significant at the 99.9% level. □

(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. ***[5 points]***

Solution. Based on this evidence, we can say with good confidence that Biden was ahead of Trump in practical and statistical terms with a high level replicability. □

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

2 Calculation (B)

[205 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

Table 2: *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

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 2 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 2 in all relevant ways}

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

(a) ***[20 total points for this sub-problem]*** Descriptively summarize (in words and numbers)

[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.

the apparent effects of aspirin on mortality in Table 2. ***[5 points]*** Do the differences observed in the table seem large to You in practical terms? ***[5 points]*** Does it look like aspirin may be beneficial? Explain briefly. ***[5 points]*** Identify the single most unusual feature of the data in Table 2. ***[5 points]***

Solution. Based on the table for all studies, except for one, that there seems to be an improvement on mortality rates through the treatment of taking aspirin. Specifically it seems that one's mortality rates goes down about 2 percentage points. On the other hand there is one saying it goes about about a percent. Overall though when treating all these studies as the same, and ignoring some details, overall the improvement doesn't seem too high, about a percent, but not Statistically Significant.
So, I would say that this difference, is large in practical terms, overall or just considering the 5 positive studies, since I think anyone would be willing to decrease their mortality rate by 1-2 percent if all one had to was to take some aspirin which doesn't have any strong side effects and is relatively affordable.
It seems like aspirin may be beneficial, the doubts come from the fact that this is not Statistically Significant so it may be that this data isn't actually useful for drawing any conclusions based on the Devil's Advocate argument discussed in class.
The most unusual feature would definitely be that the largest study on this list is showing that aspirin may be ineffective. It does suffer though, like the other one's, in not being Statistically Significant though. □

(b) ***[10 total points for this sub-problem]*** When You're comparing studies in a meta-analysis, a phenomenon called *between-study 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. A naive analysis of the data in Table 2 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 2). 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 2 provides strong evidence of between-study heterogeneity, so that naive pooling looks like a bad idea with this data set. Can You think of a medical reason why the results across the studies are so different? Explain briefly. ***[10 points]***

Solution. The biggest indicator of between-study heterogeneity is the mortality rate of the control group. We see the range of mortality rates ranges from 8.30 all the way to 14.82 percent. Keep in mind this is the group that is supposed to be receiving a placebo for a year and already then each study has very different mortality rates for this group.
This may be due to these studies having different patient cohorts from being chosen locally. Which leads to the patients themselves having large differences from other patients in the other studies. □

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 (17) 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 2, with substantial between-study heterogeneity, is as follows: for $(i = 1, \ldots, k)$,

$$\begin{aligned} (\mu\, \sigma \,|\, \mathcal{B}) &\sim p(\mu\, \sigma \,|\, \mathcal{B}) \\ (\theta_i \,|\, \mu\, \sigma\, \mathbb{N}\, \mathcal{B}) &\overset{\text{IID}}{\sim} N(\mu, \sigma^2) \\ (y_i \,|\, \theta_i\, V_i\, \mathcal{B}) &\overset{\text{I}}{\sim} N(\theta_i, V_i)\,. \end{aligned} \tag{17}$$

This is our first example of a *Bayesian hierarchical model* with more than two levels in the hierarchy: the data set summarized in Table 2 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 2;
- 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 2), 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 (17). This assumption, denoted by $\mathbb{N}$ 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.

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

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

for $\boldsymbol{\eta} \triangleq (\mu, \sigma)$ in model (17) is

$$\ell(\mu\,\sigma \,|\, \boldsymbol{y}\,\boldsymbol{V}\,\mathbb{N}\,\mathcal{B}) = \prod_{i=1}^{k} \frac{1}{\sqrt{V_i + \sigma^2}} \exp\left[-\frac{1}{2}\,\frac{(y_i - \mu)^2}{V_i + \sigma^2}\right] , \tag{18}$$

leading to the log-likelihood function

$$\ell\ell(\mu\,\sigma \,|\, \boldsymbol{y}\,\boldsymbol{V}\,\mathbb{N}\,\mathcal{B}) = -\frac{1}{2}\sum_{i=1}^{k}\left[\log(V_i + \sigma^2) + \frac{(y_i - \mu)^2}{V_i + \sigma^2}\right] . \tag{19}$$

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

$$(\hat{\boldsymbol{\eta}} \,|\, D\,\mathcal{B}) \sim N_{k_{\boldsymbol{\eta}}}\left(\boldsymbol{\eta}, \hat{I}^{-1}\right) , \tag{20}$$

in which 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.

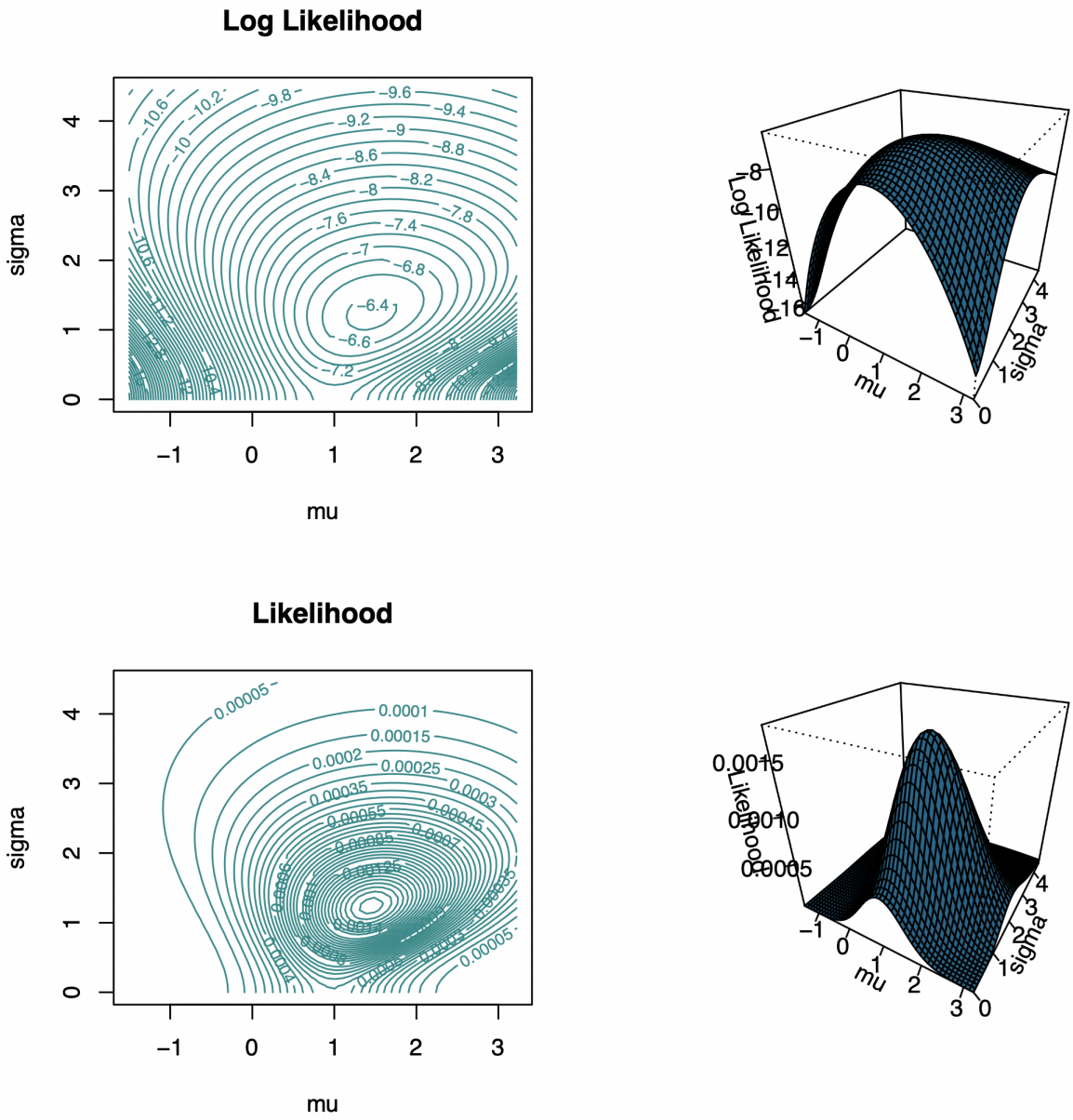
Making these plots is a bit more involved than in our previous case studies, but the basic idea is the same: in this case, we construct a two-dimensional grid in μ and σ, evaluate the ℓ and $\ell\ell$ functions on the grid, and graph them with perspective and contour plots. The main issue to settle in 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 2: 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) ***[10 total points for this sub-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

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

[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.

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 $k_{\eta} = 6$ studies. It turns out that the effective sample sizes for μ and σ are driven mainly by N and k_{η}, respectively. Do Your plots resemble the large-sample bivariate Normal shapes described above? Explain briefly. ***[5 points]***

> ***Solution.*** The log likelihood function does indeed match the what is described above since it has a concave down paraboloid while being a little bit distorted along the σ axis. The likelihood function also matches what is described because we see a peak/mountain that looks unimodal with a little bit of distortion like the log likelihood function, but not as much. □

(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. ***[5 points]***

> ***Solution.*** Yes, this is because we can clearly see from these graphs that there is no suggestion of multiple maxima, which would suggest multimodality. There is a clear "high" point in both graphs. □

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 $\ell\ell$: $\mathbb{R}^{k_{\boldsymbol{\eta}}} \to \mathbb{R}$ 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; 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) ***[45 total points for this sub-problem]*** Download the `.txt` file just mentiuoned, run my `optim` code (or an equivalent program in another language), and examine the resulting output (include this output in Your Appendix).

(i) Did the code report convergence to a (local) maximum of the log-likelihood function? ***[5 points]*** What did the MLE vector turn out to be, to 4 significant figures? ***[5 points]*** Did the maximum value of $\ell\ell$ agree with what You saw in Your plots in part (c)? ***[5 points]*** How many function evaluations did `optim` need to find the MLEs? ***[5 points]***

> ***Solution.*** Yes the code reported convergence to a local maximum of the log-likelihood function as can be seen in the appendix from the "$convergence" output.
> The MLE vector turned out to be $(1.447, 1.237)$ which is seen from "$par" output.
> The maximum value of $\ell\ell$ does agree with our plots. This is because from a quick

visual inspection of our plots, the maximum looks to be around (1.5, 1.3). Our code output matches that visual guess.
`optim` needed 43 function evaluations to find the MLE's which can be seen from the "$counts" ouput under "function".

□

(ii) Use the estimated covariance matrix of the MLEs from the `optim` output to compute estimated standard errors for $\hat{\mu}_{MLE}$ and $\hat{\sigma}_{MLE}$ (the *hint:* in the `R` code may help). ***[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 (I) made no difference or (II) was beneficial in reducing mortality. As we saw in problem 2(A)(d)(i) 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

$$\hat{\mu}_{MLE} - \Phi^{-1}(1-\alpha) \cdot \widehat{SE}\,(\hat{\mu}_{MLE}) \tag{21}$$

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 (21). Compute this LCB for $\alpha = 0.001$. ***[5 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. Completing the R code provided in order to computer the estimated standard errors we get (code in appendix),

$$\widehat{SE}(\hat{\mu}_{MLE}) = 0.8394 \qquad\qquad \widehat{SE}(\hat{\sigma}_{MLE}) = 0.6791$$

The LCB for $\alpha = 0.001$ works out to be (using R with the code in appendix) -1.147494. Based on this, we are not confident that aspirin would indeed reduce mortality of heart-attack patients in $\mathcal{P}$. This is because as we see our lower bound is well into the negative region. As discussed in class, under the Devil's Advocate arguement we see the possibility of apsirin doing nothing ($\mu = 0$) is set inside our confidence interval, making it not Statistically Significant. □

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 (17) is

$$\hat{\mu} = \frac{\sum_{i=1}^{k} \hat{W}_i\, y_i}{\sum_{i=1}^{k} \hat{W}_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \hat{W}_i^2\, [(y_i - \hat{\mu})^2 - V_i]}{\sum_{i=1}^{k} \hat{W}_i^2}\,, \quad \text{in which} \quad \hat{W}_i = \frac{1}{V_i + \hat{\sigma}^2}\,. \tag{22}$$

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 $\hat{\mu}_{MLE}$:

$$\widehat{SE}\left(\hat{\mu}_{MLE}\right) = \left[\sum_{i=1}^{k} \frac{1}{V_i + \hat{\sigma}^2_{MLE}}\right]^{-\frac{1}{2}} . \tag{23}$$

(e) ***[10 total points for this sub-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? ***[5 points]***

> ***Solution.*** In order to achieve convergence, 35 iterations were needed. This is seen under the output for "$m" which is attached in the code appendix.
> The clock time was roughly about 0.01 seconds as seen in the output. □

(ii) Your execution of the code should have produced the following results: $\hat{\mu}_{MLE} \doteq 1.447$, with an approximate estimated standard error of $\widehat{SE}\left(\hat{\mu}_{MLE}\right) \doteq 0.8089$, and $(\hat{\sigma}_{MLE},\, \hat{\sigma}^2_{MLE}) \doteq (1.237, 1.531)$. Bearing in mind (from Table 2) 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? 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. ***[5 points]***

> ***Solution.*** This is definitely large in practical terms, considering the fact that we are talking about life or death and the cost/availability of aspirin (cheap and everywhere basically). Implanting the regular consumption of aspirin in one's life, for a roughly 1.5 percent decrease in mortality rate, isn't that difficult. Therefore making it practically significant. Another way to see it, is through,
>
> $$\frac{9.55 - 11}{11} = -0.1318$$
>
> which means it decreases one's mortality rate by about 13 percent! Which is practical in medical terms.
> The amount of between-study heterogeneity corresponding to an SD of 1.24% is definitely large in practical terms. This is seen through the large range in mortality rates for the control group between all these studies. Since the control group is being given a placebo, meaning there is no treatment, yet we see substantial differences in their mortality, and our SD of 1.24% captures this. □

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 (17) given the data, and also given μ and σ, are as follows:

$$(\theta_i \,|\, y_i \,\mu \,\sigma \,\mathbb{N} \,\mathcal{B}) \overset{\mathrm{I}}{\sim} N\left[\theta_i^*, V_i(1-B_i)\right] , \quad \text{with} \quad \theta_i^* = (1-B_i)\, y_i + B_i\, \mu \quad \text{and} \quad B_i = \frac{V_i}{V_i + \sigma^2} . \tag{24}$$

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 (17) 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

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2} \quad \text{and} \quad \hat{\theta}_i = (1 - \hat{B}_i)\, y_i + \hat{B}_i\, \hat{\mu} \,, \tag{25}$$

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

$$\widehat{SE}\left(\hat{\theta}_i\right) = \sqrt{V_i \,(1 - \hat{B}_i)} \,. \tag{26}$$

(f) ***[30 total points for this sub-problem]*** Use the output from Your previous code execution in part (e) to complete Table 3, and examine the results. 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.

(i) You can see in equation (25) 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. Well to see this we have to recall the equation for the weight $\hat{B}_i$ is deter-

mined,

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}$$

meaning if a sample has a low V_i the value of its $\hat{B}_i$ will also be low (because the numerator would be small). We see from the table that study 6 has the lowest V_i, which is due from it having the largest sample size, so therefore it must have the lowest $\hat{B}_i$ weight, even though its y_i is the most different from the rest. □

(ii) Compare the p_i and $\hat{W}_i^*$ columns in Table 3. 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. Like before, this can explained by seeing how we determined $\hat{W}_i{}^*$ which we know from (22) to be calculated by,

$$\hat{W}_i{}^* = \frac{1}{V_i + \hat{\sigma}^2}.$$

So, while it is true that study 6 has the smallest V_i, it doesn't change the fact that σ^2 is large, which will make the result of the fraction small. This is because σ^2 represents the between-study heterogeneity, which is high in this meta-analysis, and study 6 actually plays a big role in making it large. □

(iii) Compute the unweighted average of the $\hat{\theta}_i$ values in Table 3. 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]***

Solution. Looking at the output of our code in the appendix, under "$theta.hat" we see these values relate to the maximum-likelihood estimation through the fact that,

$$\frac{1}{k}\sum_{i=1}^{k} \hat{\theta}_i = \hat{\mu}.$$

This is because the θ_i values are drawn from a normal curve with a mean μ, which we see in the middle line of equation (17). So for this line to be consistent and true, the relation above must be true. □

In the rest of this problem You'll perform a Bayesian analysis of the data in Table 2. Looking back at equation (17), 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 2 was the first of its kind, so we want to build a low-information (LI, otherwise known as *diffuse*) prior. There is no conjugate prior for this situation; we need to use MCMC to quantify the posterior.

It turns out that there is 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),

Table 3: *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	1.92	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	1.09
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	1.69	1.04
6	4524	0.418	0.427	0.280	0.346	−1.15	−0.251	0.728

so let's use a prior of the form $p(\mu\,\sigma\,|\,\mathcal{B}) = p(\mu\,|\,\mathcal{B}) \cdot p(\sigma\,|\,\mathcal{B})$. There are a number of ways to make this prior diffuse; research has shown two things:

- The posterior is insensitive to the precise details specifying $p(\mu\,|\,\mathcal{B})$ 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 $(\mu\,|\,\mathcal{B}) \sim \text{Uniform}(A, B)$, where A and B are chosen to avoid inappropriate truncation of the posterior; and
- Care ***is*** required in specifying $p(\sigma\,|\,\mathcal{B})$ 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 a diffuse prior on σ is $(\sigma\,|\,\mathcal{B}) \sim \text{Uniform}(0, C)$, where 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)`

After some experimentation I chose $(A, B, C) = (-2, 5, 6)$ in the prior specification. Download the `.txt` file just mentioned, run parts (0)–(11) 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) ***[60 total points for this sub-problem]*** Use the output from Your MCMC code execution to complete Table 4 by filling in the blank entries; answering the questions below will also involve extracting additional numbers from the output.

> The plots and the code output during my execution will be in the code appendix.

(i) 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 (17), has shown that Bayes and ML findings will either be similar (when k is large) or the ML approach will often underestimate uncertainty when

Table 4: *Maximum-likelihood and Bayesian results in the aspirin meta-analysis; — means that results with the indicated method for the indicated quantity are not available.*

	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	2.096	1.319
θ_2	1.943	—	0.8995	2.042	1.130
θ_3	1.533	—	1.094	1.592	1.542
θ_4	1.841	—	0.9941	1.984	1.315
θ_5	1.692	—	1.049	1.812	1.431
θ_6	−0.2514	—	0.7278	−0.4327	0.9425

it differs from Bayes. Does the second of those two possibilities appear to have happened here? Explain briefly. ***[5 points]***

Solution. Well from our table we see that the pertenage points of mortality reduction for the posterior mean μ is about 1.502 and under ML it is about 1.447. Which are somehwat close to each other.
The posterior standard deviation for μ is about 1.056, and the under ML, information based standard error is 0.8394 and the emperical bayes is 0.8089. Here we see a bigger difference, under ML we have less uncertainty compared to the posterior uncertainty. We see from our comparision that the second case is happening here, since our k is only 6, and the ML approach is underestimating our uncertainty compared to Bayes. Since our uncertainty under Bayes is about 1.056 which is considerably larger than 0.8394 and 0.8089. Meaning Bayes provides a better SD than ML since it is an underestimate due to the small k.

□

(ii) 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. ***[5 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. ***[5 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. We see the posterior mean for σ to be 1.896 and the ML mean for σ is 1.237. We see that the posterior mean for σ is about 50 percent larger than the ML mean for σ. It is clear from this that it certaintly matters which one we choose to report, due to the substantial difference.
From the code provided we see our 99.9% Bayesian posterior interval for σ is:

$$(0.002, 5.94)$$

Now calculating the large-sample-approximate 99.9% CI for σ from the maximum-likelihood is simply,

$$1.237 \pm \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) 0.6791 = 1.237 \pm 3.29 \cdot 0.6791 = (-1.0, 3.5)$$

which is well into the negative region.
We wouldn't agree with her here because we can see that 0 is not in our 99.9% Bayesian confidence interval, meaning that σ is statistically significant from 0 at this level of confidence based on the data.

□

(iii) Show (by extracting the relevant number from Your output) that, conditional on model (17) 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%. ***[5 points]*** Is this standard of envidence 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. We see that the probability of μ being positive (beneficial) when conditioning on the data, the low information prior, model 17, and the background, we get it to be about 0.9332 or about 93% as desired.
Under today's standards of replicability no, but personally I would recommend it if one is willing, but only to the people like the ones in studies 1-5. This is because for the people of study 6, aspirin was not found to be helpeful at all. The reason I would recommend it, if people woudld like, is because the cost and side effects of taking aspirin are not large or worrying, so for the potential gain that may be recieved (not dying) it may be worth it. I definitely wouldn't strongly advise it or recommend to make it some sort of standard treatment based on this though. □

(h) ***[20 total points for this sub-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* model for $(i = 1, \ldots, k)$:

$$\begin{aligned} (\mu \,|\, \mathcal{B}) &\sim p(\mu \,|\, \mathcal{B}) \\ (y_i \,|\, \mu\, V_i\, \mathcal{B}) &\overset{\text{IID}}{\sim} N(\mu, V_i)\,. \end{aligned} \tag{27}$$

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*.

Table 5: *DIC comparison of the fixed effects and random effects models in the aspirin meta-analysis.*

Model	Mean Deviance	Complexity Penalty	*DIC*
Fixed Effects	27.1	1.0	28.1
Random Effects	21.6	4.1	25.7

The plot and code output will be in the code appendix.

(i) By examining the random effects model equations (17), briefly explain why the fixed effects model in (27) is a special case of (17) in which it's assumed that $\sigma = 0$. ***[5 points]***

Solution. Well σ is representing our between-study heterogeneity, so if σ were to be 0, our middle layer in modle (17) will dissapear since all the θ_i's would be deterministically equal to μ (because a Normal distribution with mean μ and variance 0 would be a point mass at μ). Meaning we would only need a prior for μ, giving us model (27). Therefore model (27) is a special case of model (17) where $\sigma = 0$. □

(ii) Run the final block of code (section (12) in the `rjags` code file) to get *DIC* values for the fixed effects and random effects models. Use your output to fill in the missing (blank) entries in Table 5. 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? Does this agree with your conclusions about between-study heterogeneity in the earlier parts of this problem? Explain briefly. ***[15 points]***

Solution. It is clear that the random effects models is better than fixed effect models. It is difficult to say if this differene is enough though to *strongly* suggest it over fixed effects. What we can say is that random effect models is better here, but we can't say how much better based on this difference. This does agree with our conclusions of between-study heterogeneity since using the model that takes it into account turned out to be better than assuming there to be none. So our suspicion of between-study heterogeneity turned out to be correct. □

CODE FOR PROBLEMS Apart from the the code explicitly given to use for certain problems, a lot of this code taken or influenced from the office hours given by professor Draper.

START ==== 2A − di(Method1) ==========================

```
n.1 <- 659 #Biden
n.2 <- 554 #Trump
n.3 <- 52 #Other

n <- n.1 + n.2 + n.3 #Total sample

                     #Proportions
theta.1 <- n.1 / n   #Biden
theta.2 <- n.2 / n   #Trump
theta.3 <- n.3 / n   #Other

gamma <- theta.1 - theta.2

#Variances of theta.1, theta.2 and theta.3
SE.hat.theta.1 = sqrt((theta.1 * (1 - theta.1) )/n )
SE.hat.theta.2 = sqrt((theta.2 * (1 - theta.2) )/n )
SE.hat.theta.3 = sqrt((theta.3 * (1 - theta.3) )/n )

estimated.variance.gamma <- (theta.1 * (1-theta.1)/n) +
    (theta.2 * (1-theta.2)/n) + 2*( (theta.1 * theta.2)/ n)

SE.hat.gamma <- sqrt(estimated.variance.gamma)

print(estimated.variance.gamma)
print(SE.hat.gamma)

print(theta.1)
print(theta.2)
print(theta.3)
```

END ==== 2A − di(Method1) ==========================

START ======== 2A − f(ii) ==========================

```
options(scipen=999)
rdirichlet <- function( M, alpha ) {
  k <- length( alpha )
  theta.star <- matrix( NA, M, k )
  for ( j in 1:k ) {
    theta.star[ , j ] <- rgamma( M, alpha[ j ], 1 )
```

```
  }
  theta.star <- theta.star / apply( theta.star, 1, sum )
  return( theta.star )
}

epsilon <- 0.01
k <- 3
alpha.prior <- rep(epsilon, k)

n.1 <- 659 #Biden
n.2 <- 554 #Trump
n.3 <- 52 #Other

n <- c(n.1, n.2 , n.3)

alpha.posterior <- alpha.prior + n

M.0 <- 100000 #pilot M
str(
  initial.MC.data.set <- rdirichlet(M.0, alpha.posterior)
)

initial.MC.posterior.mean.estimates <-apply(initial.MC.data.set, 2, mean)

initial.MC.posterior.sd.estimates <- apply(initial.MC.data.set, 2, sd)

theta.1.star <- initial.MC.data.set[, 1]
theta.2.star <- initial.MC.data.set[, 2]
theta.3.star <- initial.MC.data.set[, 3]

gamma.star <- theta.1.star - theta.2.star

initial.MC.data.set.gamma.star <- cbind(theta.1.star,
    theta.2.star, theta.3.star, gamma.star)


#Standard error of each theta_i
mcse.theta.1.bar.star <- sd(theta.1.star)/ sqrt(M.0)
mcse.theta.2.bar.star <- sd(theta.2.star)/ sqrt(M.0)
mcse.theta.3.bar.star <- sd(theta.3.star)/ sqrt(M.0)

gamma.star.mean <- mean(gamma.star)

#Standard Error of gamma
mcse.gmma.bar.star <- sd(gamma.star) / sqrt(M.0)

#The needed M which we use next
needed.M.0 <- ceiling((sd(gamma.star)/ 0.00005)^2)
```

START OF USING UPDATED M VALUE ==========

```
M.0 <- needed.M.0 #updated M
str(
  final.MC.data.set <- rdirichlet(M.0, alpha.posterior)
)

final.MC.posterior.mean.estimates <-apply(final.MC.data.set, 2, mean)

final.MC.posterior.sd.estimates <- apply(final.MC.data.set, 2, sd)

theta.1.star <- final.MC.data.set[, 1]
theta.2.star <- final.MC.data.set[, 2]
theta.3.star <- final.MC.data.set[, 3]

gamma.star <- theta.1.star - theta.2.star

final.MC.data.set.gamma.star <- cbind(theta.1.star,
    theta.2.star, theta.3.star, gamma.star)


mcse.theta.1.bar.star <- sd(theta.1.star)/ sqrt(M.0)
mcse.theta.2.bar.star <- sd(theta.2.star)/ sqrt(M.0)
mcse.theta.3.bar.star <- sd(theta.3.star)/ sqrt(M.0)

mcse.gmma.bar.star <- sd(gamma.star) / sqrt(M.0)
```

END ======= 2A − f(ii) ======================

START ===== 2A − f(v) ========

```
#This is the code that gave the values,
#but it uses the code right above
print(mean(gamma.star > 0))
print(sd(gamma.star > 0)/ sqrt(M.0))
```

END ===== 2A − f(v) ========

START ==== 2B − c(i)&c(ii) ======================

```
#Used R file provided to us on canvas to generate graphs
```

END ====== 2B − c(i)&c(ii) ======================

START ===== 2B − d(i) ==============================

```
#This is code given to us, but I deleted all the comments due to length
# enter the data:

y <- c( 2.77, 2.50, 1.84, 2.56, 2.32, -1.15 )
V <- c( 1.65, 1.31, 2.34, 1.67, 1.98, 0.90 )^2

aspirin.mortality.log.likelihood.for.optim <- function( eta, y, V ) {
  mu <- eta[ 1 ]
  sigma <- eta[ 2 ]
  ll <- ( -1 / 2 ) * sum( log( V + sigma^2 ) + ( y - mu )^2 /
                                ( V + sigma^2 ) )
  return( ll )
}

eta.initial.values <- c( 1.5, 1.25 )
print( ml.results.1 <- optim( eta.initial.values,
                         aspirin.mortality.log.likelihood.for.optim,
                         y = y, V = V,
                         hessian = T, control = list( fnscale = -1 ) ) )

print( maximum.likelihood.covariance.matrix <-
          solve( - ml.results.1$hessian ) )

print(
    maximum.likelihood.estimated.standard.errors <-
        diag(maximum.likelihood.covariance.matrix)
    )

#OUTPUT
#$par
#[1] 1.446576 1.237251

#$value
#[1] -6.332311

#$counts
#function gradient
#       43       NA
```

```
#$convergence
#[1] 0

#$message
#NULL

#$hessian
#             [,1]       [,2]
#[1,] -1.5279769  0.5041351
#[2,]  0.5041351 -2.3347060

#            [,1]      [,2]
#[1,] 0.7046628 0.1521584
#[2,] 0.1521584 0.4611751
#[1] 0.8394419 0.6790988
```

END ===== 2B − d(i) ==========

START ====== 2B − d(ii) ==========

```
alpha <- 0.001

print(1.446576 - qnorm(1 - alpha) * 0.8394419)
```

END ===== 2B − d(ii) ==========

START ===== 2B − e(i)(ii)(iii) ==========

```
#Used the code provided from the .txt file mentioned in the
#problem statement
#When ran this is the OUTPUT
$m
[1] 35

$mu.hat
[1] 1.446869

$se.hat.mu.hat
[1] 0.8089829
```

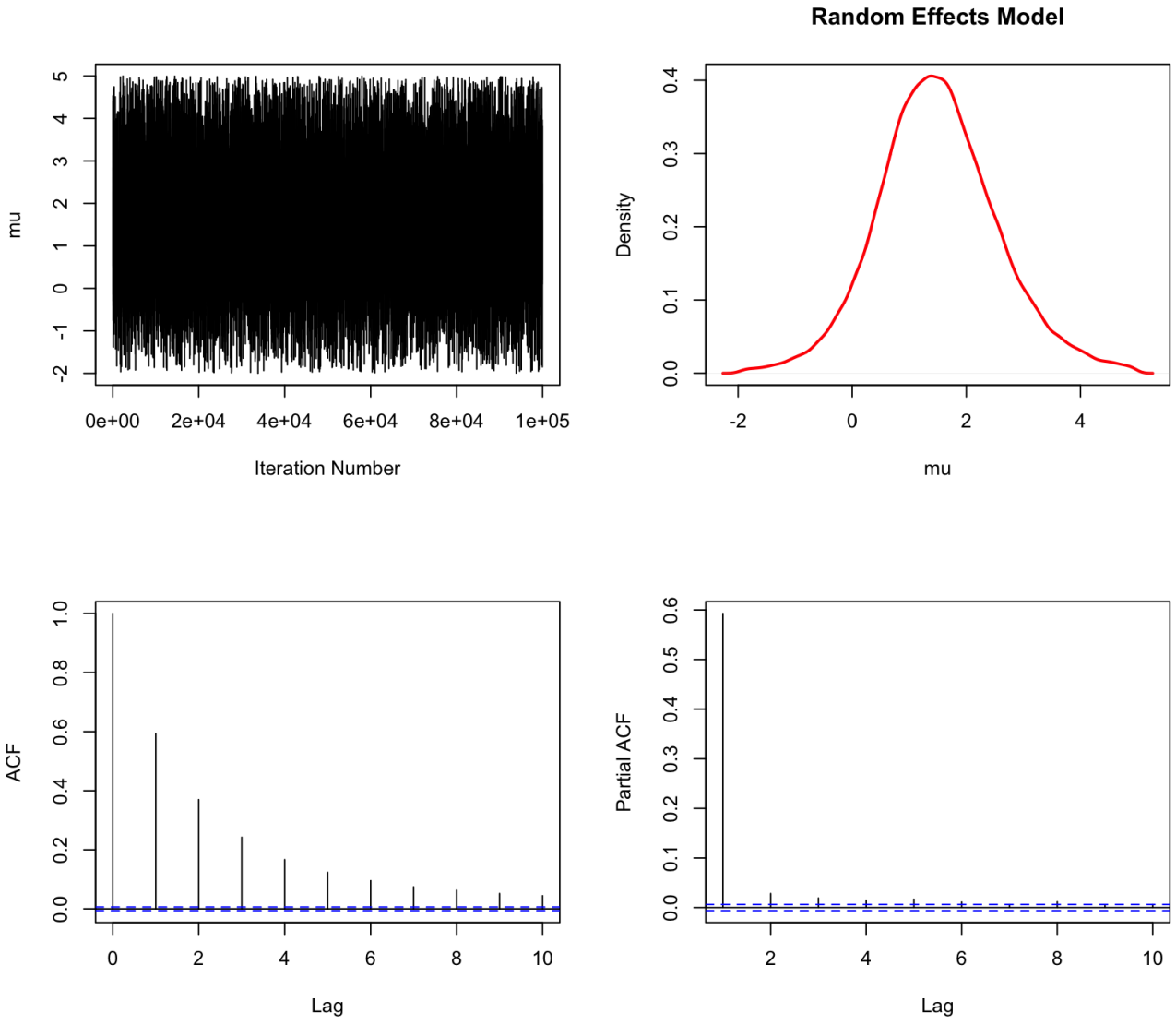
```
$sigma.squared.hat
[1] 1.530753

$sigma.hat
[1] 1.237236

$n
[1] 1239 1529  626 1682 1216 4524

$n.normalized
[1] 0.11455251 0.14136464 0.05787722 0.15551036 0.11242604 0.41826923

$W.hat
[1] 0.2351142 0.3079905 0.1427276 0.2315001 0.1834474 0.4272130

$W.hat.normalized
[1] 0.15387125 0.20156544 0.09340856 0.15150600 0.12005779 0.27959095

$B.hat
[1] 0.6400983 0.5285426 0.7815193 0.6456306 0.7191873 0.3460425

$y
[1]  2.77  2.50  1.84  2.56  2.32 -1.15

$theta.hat
[1]  1.9230658  1.9433752  1.5327602  1.8413283  1.6920550 -0.2513731

$se.hat.theta.hat
[1] 0.9898648 0.8994821 1.0937609 0.9941332 1.0492369 0.7278087

   user  system elapsed
  0.011   0.000   0.011
```

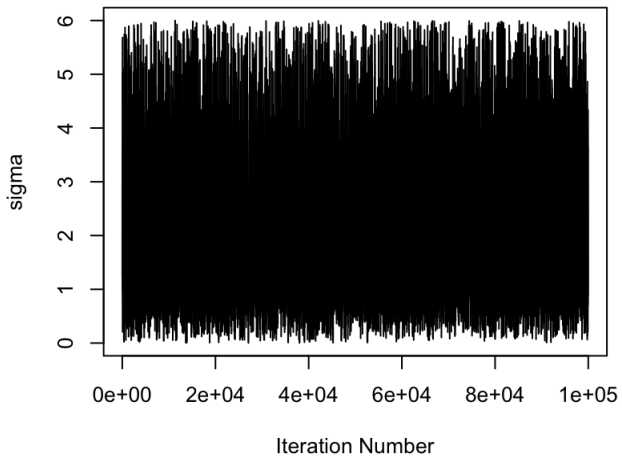
END ===== 2B − e(i)(ii)(iii) ==========

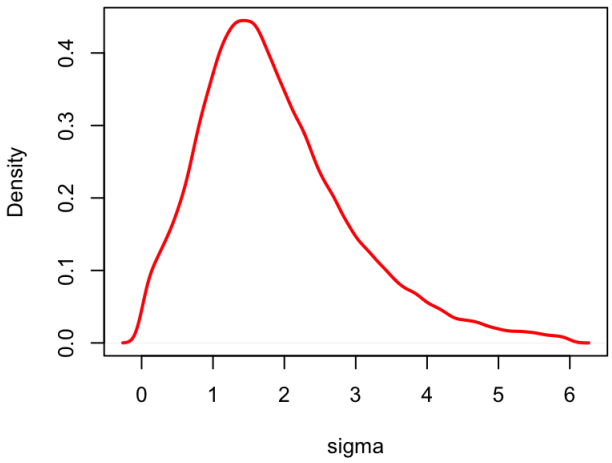
START ======= 2B − g(i)(ii)(iii) ==========

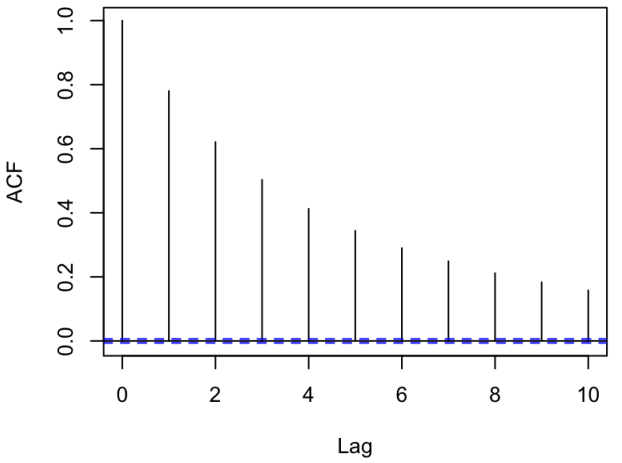
```
#Code used was provided, I didn't modify it other than to print values
#so I won't be pasting it here to save space
#CODE OUTPUT
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
```

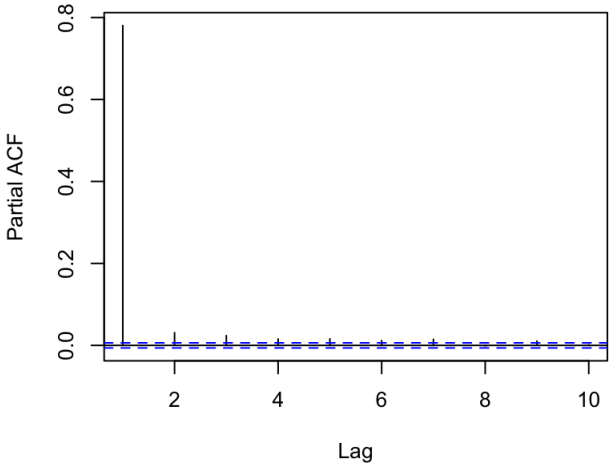
```
Graph information:
   Observed stochastic nodes: 6
   Unobserved stochastic nodes: 8
   Total graph size: 30

Initializing model

  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
  |**************************************************| 100%
  |**************************************************| 100%
[1] 0.567431
[1] 0.006355817
                          0.05%     99.95%
1.8874320 1.0879348 0.0105058 5.9327306
[1] 0.7766925
[1] 0.009704146
 'mcarray' num [1:6, 1:100000, 1] 2.95 1.76 3.77 3.35 3.52 ...
 - attr(*, "varname")= chr "theta"
 - attr(*, "type")= chr "trace"
 - attr(*, "iterations")= Named num [1:3] 2001 102000 1
  ..- attr(*, "names")= chr [1:3] "start" "end" "thin"
[1]  2.0957896  2.0390408  1.5898098  1.9838042  1.8203684 -0.4211204
[1] 1.318550 1.129905 1.535767 1.316264 1.431770 0.951049
[1] 0.93344
```

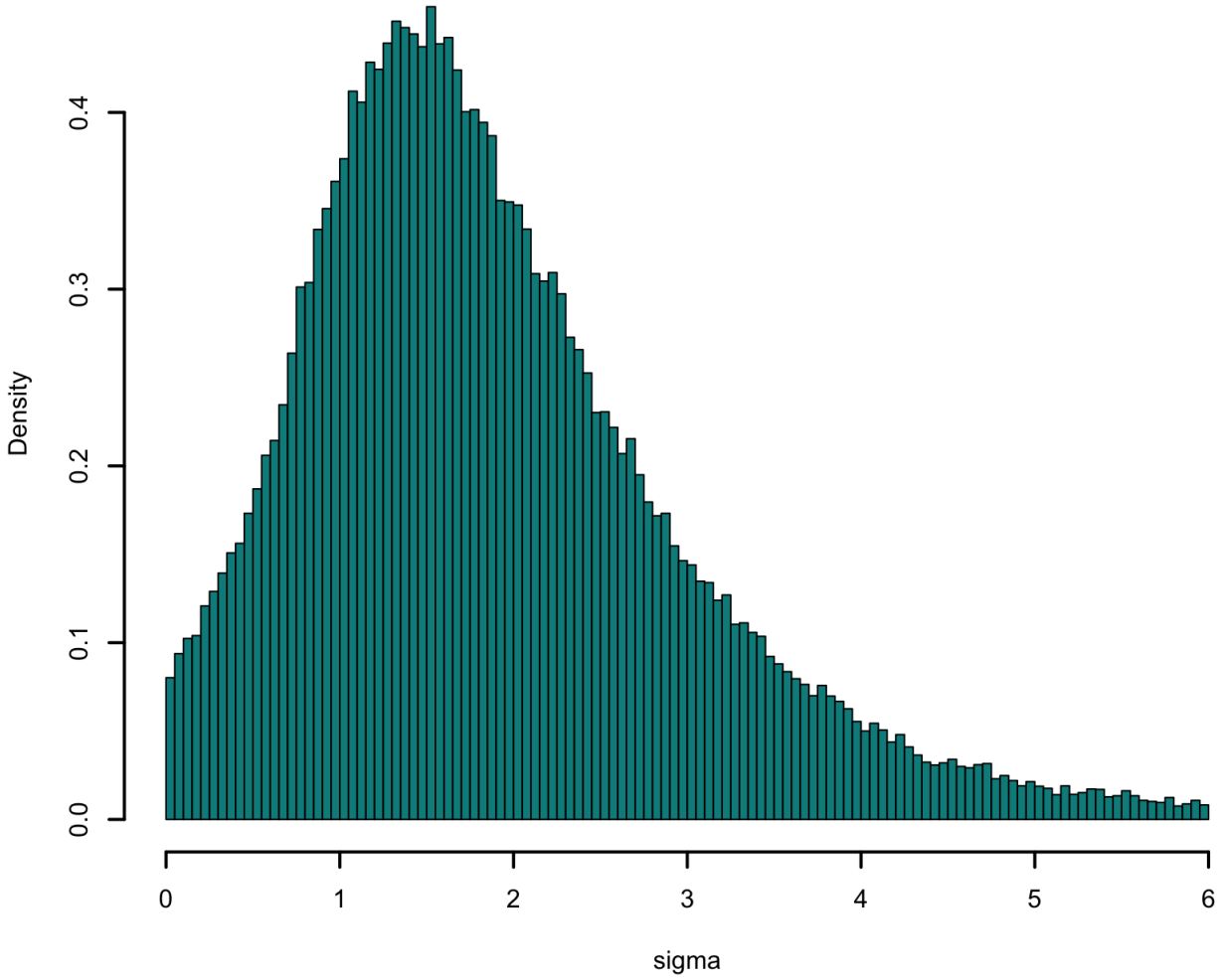

Random Effects Model
mu
Iteration Number
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Density
mu
ACF
Lag
Partial ACF
Lag

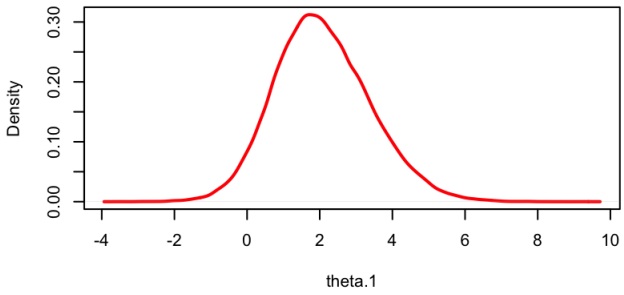

sigma
0
1
2
3
4
5
6
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Iteration Number

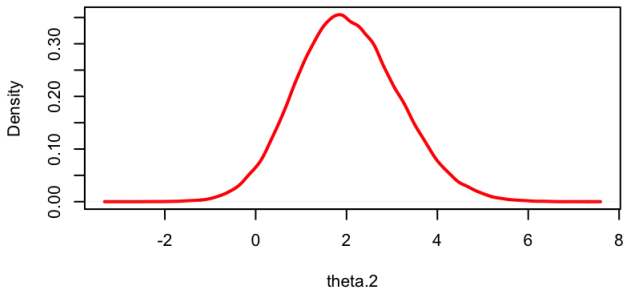

Density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
sigma

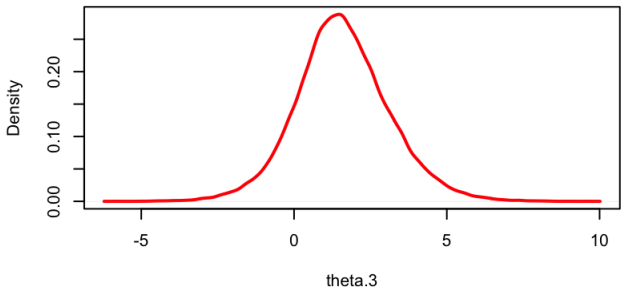

ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Lag

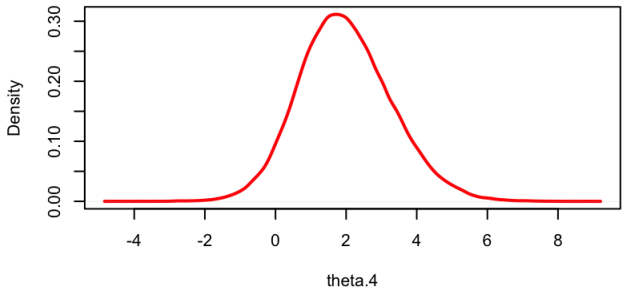

Partial ACF
0.0
0.2
0.4
0.6
0.8
2
4
6
8
10
Lag

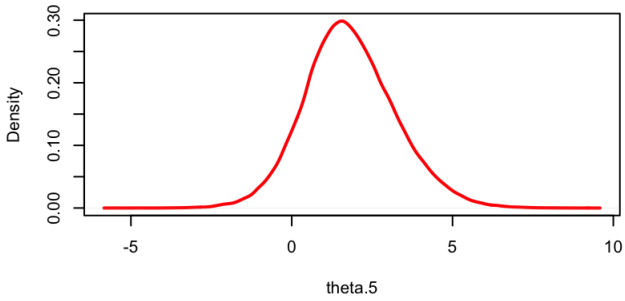

Density
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
6
sigma

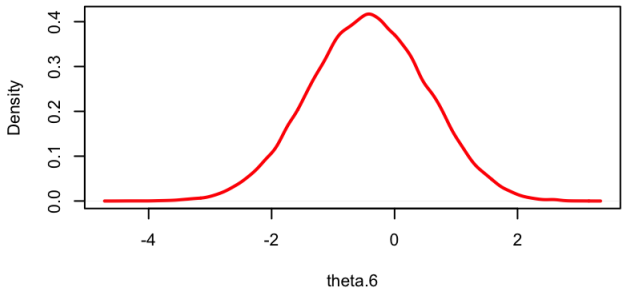










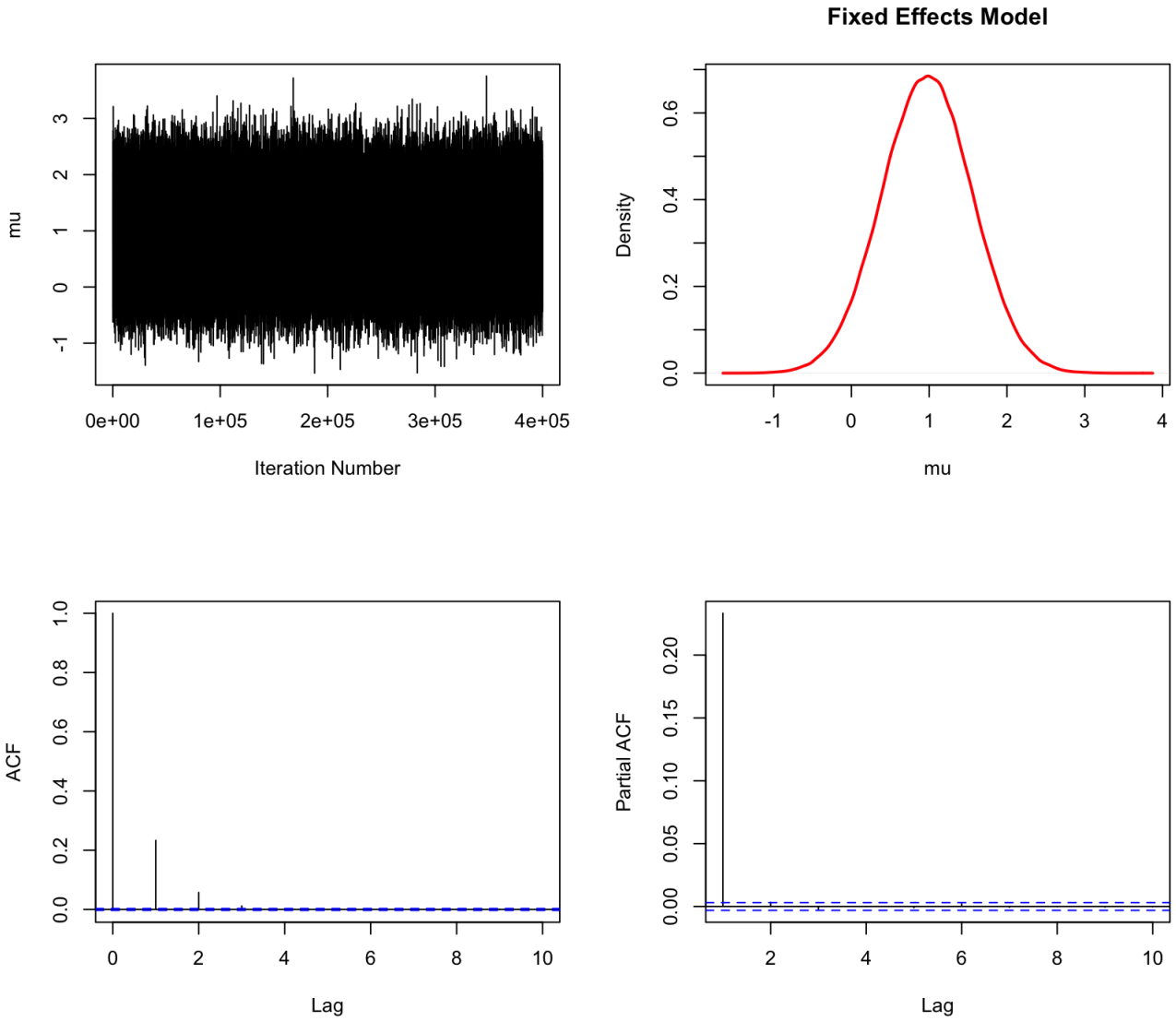


***END* ======= $2B - g(i)(ii)(iii)$ ==========**

***START* ====== $2B - h(i)(ii)$ ==========**

```
#Used code provided
#CODE OUTPUT
Compiling model graph
    Resolving undeclared variables
    Allocating nodes
Graph information:
    Observed stochastic nodes: 6
    Unobserved stochastic nodes: 8
```

```
   Total graph size: 30

Initializing model

  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
  |**************************************************| 100%
Mean deviance:  21.63
penalty 4.066
Penalized deviance: 25.7
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 6
   Unobserved stochastic nodes: 1
   Total graph size: 18

Initializing model

  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100%
  |**************************************************| 100%
[1] 0.2332276
[1] 0.001159899
[1] 0.95403
  |**************************************************| 100%
Mean deviance:  27.07
penalty 1
Penalized deviance: 28.07
```



$\boldsymbol{END}$ ====== $2B - h(i)(ii)$ ==========