University of California, Santa Cruz
Department of Applied Mathematics and Statistics
Baskin School of Engineering

AMS 206 (Applied Bayesian Statistics)

Take-Home Test 2 (final version)

Due date at `canvas.ucsc.edu`: by 11.59pm on **Sun 10 Mar 2019**

Here are the 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 10 points, and the calculation problem is worth 280 total points (with possible additional extra credit of up to 25 points), for a total of 340 points**.

The right answer with no reasoning to support it, or incorrect reasoning, will get **half credit**, so try to make a serious effort on each part of each problem (this will ensure you at least half credit). In an AMS graduate class I taught in 2012, on a take-home test like this one there were 15 true/false questions, worth a total of 150 points; one student got a score of 92 out of 150 (61%, a D−, in a graduate class where B− is the lowest passing grade) on that part of the test, for repeatedly answering just "true" or "false" with no explanation. Don't let that happen to you.

On non-extra-credit problems, I mentally start everybody out at −0 (i.e., with a perfect score), and then you accumulate negative points for incorrect answers and/or reasoning, or parts of problems left blank. On extra-credit problems, the usual outcome is that you go forward (in the sense that your overall score goes up) or you at least stay level, but please note that it's also possible to go backwards on such problems (e.g., if you accumulate +3 for part of an extra-credit problem but −4 for the rest of it, for saying or doing something egregiously wrong).

This test is to be entirely your own efforts; do not collaborate with anyone or get help from anyone but me or our TA (René Gutierrez). 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 (because people from your cultural background routinely do this, or out of pity, or kindness, or whatever motive you may believe you have; it doesn't matter), you're just as guilty of illegal collaboration as the person who took your solution from you, and both of you will receive the same penalty. This sort of thing is necessary on behalf of the many people who do not cheat, to ensure that their scores are meaningfully earned. In the AMS graduate class in 2012 mentioned above, five people failed the class because of illegal collaboration; don't let that happen to you.

In class I've demonstrated numerical work in `R`; you can (of course) make the calculations and plots requested in the problems below in any environment you prefer (e.g., `Matlab`, `Python`, ...).

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 better give you part credit. 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 you can say something like the following:

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

1 True/False

[60 total points: 10 points each] For each statement below, say whether it's true or false; if true without further assumptions, briefly explain why it's true (and — *extra credit* — 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 $(Y_i \,|\, \boldsymbol{\theta}\, \mathcal{B}) \overset{\text{IID}}{\sim} p(y_i \,|\, \boldsymbol{\theta}\, \mathcal{B})$ for $i = 1, \dots, n$, where the Y_i are univariate real values, $\boldsymbol{\theta}$ is a parameter vector of length $1 \le k < \infty$ 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 diffuse (low information content), but the theorem does not provide guidance on how large n needs to be for its conclusion to hold in any specific sampling model.

(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 page 2 of the document camera notes from 24 Jan 2019) — 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.

(C) Being able to express Your sampling distribution as a member of the Exponential Family is helpful, because

- You can then readily identify a set of 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.

(D) When the sampling model is a regular parametric family $p(\boldsymbol{Y} \,|\, \boldsymbol{\theta}\, \mathcal{B})$, where $\boldsymbol{\theta}$ is a vector of length $1 < k < \infty$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, then 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 $O\big(\frac{1}{n^2}\big)$.

(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).

(F) When Your sampling model has n observations and a single parameter θ (so that $k = 1$), if the sampling model is regular (i.e., if the range of possible data values doesn't depend on θ), 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 $O\big(\frac{1}{n}\big)$.

2 Calculation

(A) *[95 total points, plus a total of 25 possible extra-credit points]* (Based on a problem in Gelman et al. (2014)) In late October 1988, a survey was conducted on behalf of *CBS News* of n = 1,447 adults aged 18+ in the United States, to ask about their preferences in the upcoming presidential election. Out of the 1,447 people in the sample, $n_1 = 727$ supported George H.W. Bush, $n_2 = 583$ supported Michael Dukakis, and $n_3 = 137$ 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 of voting age in the U.S. in October 1988}. There were about 245 million Americans in 1988, of whom about 74% were 18 or older, so $\mathcal{P}$ had about 181 million people in it; the total sample size of n = 1,447 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 1,447 people in the survey is exchangeable — the only appropriate sampling distribution for the 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 AMS 131 notes, or DeGroot and Schervish (2012), to renew Your acquaintance with the Multinomial). Suppose that a population of interest contains items of $k \geq 2$ types (in the example here: people who support {Bush, Dukakis, 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 *CBS News* 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 keeps track of all of the counts. In this situation people say that $\boldsymbol{N}$ follows the Multinomial distribution

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

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

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

with the further restriction that $0 \le n_j \le n$ (for all $j = 1, \dots, k$). The main scientific and political interest in this problem focuses on $\gamma = (\theta_1 - \theta_2)$, the margin by which Bush was leading Dukakis on the day of the survey.

(a) *[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

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

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]*.

(b) *[15 total points for this sub-problem, plus up to 10 possible extra credit points]* Returning now to the general Multinomial setting, briefly explain why the likelihood function for $\boldsymbol{\theta}$ given $\boldsymbol{N}$ and $\mathcal{B}$ is

$$\ell(\boldsymbol{\theta} \,|\, \boldsymbol{N} \, \mathcal{B}) = c \prod_{j=1}^{k} \theta_j^{n_j} \,, \tag{3}$$

leading to the log-likelihood function (ignoring the irrelevant constant)

$$\ell\ell(\boldsymbol{\theta} \,|\, \boldsymbol{N} \, \mathcal{B}) = \sum_{j=1}^{k} n_j \, \log \theta_j \,. \tag{4}$$

[5 points]. 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 $\ell\ell(\boldsymbol{\theta} \,|\, \boldsymbol{N} \, \mathcal{B})$ 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 $\sum_{j=1}^{k} \theta_j = 1$. There are thus two ways forward to compute the MLE:

(i) Solve the constrained optimization problem directly with *Lagrange multipliers (Extra credit [5 points]: do this)*, or

(ii) build the constraint directly into the likelihood function: define

$$\ell(\theta_1, \dots, \theta_{k-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = c \left(\prod_{j=1}^{k-1} \theta_j^{n_j} \right) \left(1 - \sum_{j=1}^{k-1} \theta_j \right)^{n_k} , \tag{5}$$

from which (ignoring the irrelevant constant)

$$\ell\ell(\theta_1, \dots, \theta_{k-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = \sum_{j=1}^{k-1} n_j \, \log \theta_j + n_k \log \left(1 - \sum_{j=1}^{k-1} \theta_j \right) . \tag{6}$$

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

$$\frac{\partial}{\partial\, \theta_j} \, \ell\ell(\theta_1, \dots, \theta_{k-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = \frac{n_j}{\theta_j} - \frac{n_k}{1 - \sum_{i=1}^{k-1} \theta_i} \tag{7}$$

[5 points]. The MLE for $(\theta_1, \dots, \theta_{k-1})$ may now be found by setting $\frac{\partial}{\partial\, \theta_j} \ell\ell(\theta_1, \dots, \theta_{k-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = 0$ for $j = 1, \dots, (k-1)$ and solving the resulting system of $(k-1)$ equations in $(k-1)$ unknowns *(Extra credit [5 points]: do this for general k)*, but that gets quite messy; let's just do it for $k = 3$, which is all we need for the CBS survey anyway. Solve the two equations

$$\left\{ \frac{n_1}{\theta_1} - \frac{n_3}{1 - \theta_1 - \theta_2} = 0, \quad \frac{n_2}{\theta_2} - \frac{n_3}{1 - \theta_1 - \theta_2} = 0 \right\} \tag{8}$$

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

$$\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3 \right) = \left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n} \right) \tag{9}$$

[5 points]. (The result for general k, of course, is that $\hat{\boldsymbol{\theta}} = \frac{1}{n} \boldsymbol{N}$.)

(c) *[20 total points for this sub-problem, plus up to 5 possible extra credit points]* It can be shown *(Extra credit [5 points]: do this for general k, by working out the negative Hessian, evaluated at the MLE, to get the information matrix $\hat{\boldsymbol{I}}$ and then inverting $\hat{\boldsymbol{I}}$)* 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

$$\hat{\boldsymbol{\Sigma}} = \left(\begin{array}{ccc} \frac{\hat{\theta}_1 (1 - \hat{\theta}_1)}{n} & -\frac{\hat{\theta}_1 \, \hat{\theta}_2}{n} & -\frac{\hat{\theta}_1 \, \hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1 \, \hat{\theta}_2}{n} & \frac{\hat{\theta}_2 (1 - \hat{\theta}_2)}{n} & -\frac{\hat{\theta}_2 \, \hat{\theta}_3}{n} \\ -\frac{\hat{\theta}_1 \, \hat{\theta}_3}{n} & -\frac{\hat{\theta}_2 \, \hat{\theta}_3}{n} & \frac{\hat{\theta}_3 (1 - \hat{\theta}_3)}{n} \end{array} \right) . \tag{10}$$

Explain why the form of the diagonal elements of $\hat{\boldsymbol{\Sigma}}$ makes good intuitive sense (by thinking about the corresponding results when there are only $k = 2$ outcome categories); also explain why it makes good sense that the off-diagonal elements of $\hat{\boldsymbol{\Sigma}}$ are negative *[5 points]*. 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

(i) work it out directly by thinking about the repeated-sampling variance of the difference of two (correlated) random quantities, or

(ii) use the fact (from AMS 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

$$\hat{V}(\hat{\gamma}) = \hat{V}\left(\boldsymbol{a}^T \hat{\boldsymbol{\theta}}\right) = \boldsymbol{a}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{a} \tag{11}$$

[5 points]. Finally, use Your estimated SE for $\hat{\gamma}$ to construct an approximate (large-sample) 95% confidence interval for γ *[5 points].* Was Bush ahead of Dukakis at the point when the survey was conducted by an amount that was large in practical terms? Was Bush's lead at that point statistically significant? Explain briefly. *[5 points]*

(d) *[10 total points for this sub-problem]* Looking back at equation (3), 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, with $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$ chosen so that all of the α_j are positive:

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = c \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} , \tag{12}$$

in which $\mathcal{D}$ stands for the Dirichlet prior disribution assumption, which is not part of $\mathcal{B}$. Briefly explain why this means that the conjugate updating rule is

$$\left\{ \begin{array}{ccc} (\boldsymbol{\theta} \,|\, \mathcal{D}) & \sim & \text{Dirichlet}(\boldsymbol{\alpha}) \\ (\boldsymbol{N} \,|\, \boldsymbol{\theta}\, \mathcal{B}) & \sim & \text{Multinomial}(n, \boldsymbol{\theta}) \end{array} \right\} \longrightarrow (\boldsymbol{\theta} \,|\, \boldsymbol{N}\, \mathcal{D}\, \mathcal{B}) \sim \text{Dirichlet}(\boldsymbol{\alpha} + \boldsymbol{N}) \tag{13}$$

[5 points]. Given that $\boldsymbol{N} = (n_1, \dots, n_k)$ 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-content prior, this would correspond to choosing the α_j all close to 0. *[5 points]*

(e) *[45 total points for this sub-problem, plus up to 10 possible extra credit points]* 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 (13) to do MCMC sampling: IID Monte Carlo sampling is sufficient *[5 points].* It turns out that the following is a valid way to sample a vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ 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 $(j = 1, \dots, k)$, make k independent draws g_j with draw j from the $\Gamma(\alpha_j, \beta)$ distribution; and

∗ then just normalize:

$$g_j \overset{\text{I}}{\sim} \Gamma(\alpha_j, \beta) \quad \text{and} \quad \theta_j = \frac{g_j}{\sum_{i=1}^{k} g_i} , \tag{14}$$

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

I've written an R function called `rdirichlet`, posted on the course web page, that implements this algorithm. Use my function (or an equivalent in Your favorite non-R environment) to generate M IID draws from the posterior distribution specified by model (13), using the CBS News 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 $\gamma = (\theta_1 - \theta_2)$ *[10 points]*. 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 *[5 points]*. Make graphical and numerical summaries of the posterior distributions for γ and for each of the components of $\boldsymbol{\theta}$, and compare Your posterior distribution for γ with Figure 3.2 (p. 70) from the Gelman et al. (2014) book that's available at the course web site; also compute the 95% central posterior interval for γ *[10 points]*. How do Your Bayesian answers compare with those from maximum likelihood in this problem? Explain briefly *[5 points]*. Compute a Monte Carlo estimate of $p(\gamma > 0 \,|\, \boldsymbol{N}\, \mathcal{D}\, \mathcal{B})$, which quantifies the current information about whether Bush is leading Dukakis in the population of all adult Americans, and attach a Monte Carlo standard error to Your estimate; on the basis of this Bayesian calculation, is Bush's lead statistically significant? *[5 points]*. What substantive conclusions do You draw about where the Presidential race stood in late October of 1988, on the basis of Your analysis? Explain briefly *[5 points]*. *(Extra credit [10 points]: Use **`Maple`** or some other symbolic-computing environment (or paper and pen, if You're brave) to see if You can derive a closed-form expression for $p(\gamma > 0 \,|\, \boldsymbol{N}\, \mathcal{D}\, \mathcal{B})$, and compare Your mathematical result with Your simulation-based findings; if no such expression is possible, briefly explain why not.)*

(B) *[185 total points]* (This problem looks hard just because it's long, but it's 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 this problem just to set up the relevant scientific and statistical questions.) One of the most important priorities in treating patients who have just suffered a heart attack is to prevent a second heart attack or stroke, which can occur shortly after the first attack if one or more blood clots enters the blood stream and lodges in the heart or brain. This suggests that the administration of a blood-thinning drug (which would break up blood clots and prevent their formation) right after the first attack may keep the patient from dying from another immediate attack. One such drug is a low dose (as low as 75mg) of the common pain-relief drug aspirin (the usual dose for pain is 350–650mg every four hours).

Table 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 and who were representative of such people (in their region of their country) were randomized, 615 to a *treatment group* that received a low-dose aspirin each day for three months, and 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 $y_1 = 2.77$ 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)

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	
	Number	Mortality	Number	Mortality	Difference	$\sqrt{V_i} = \widehat{SE}$
Study (i)	of Patients	Rate (%)	of Patients	Rate (%)	(y_i) (%)	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

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 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.

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

(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 is a bad idea with this data set. Can You think of a medical reason why the results across studies are so different? Explain briefly. *[10 points]*

A standard Bayesian model for meta-analytic data with substantial between-study heterogeneity is as follows:

$$
\begin{aligned}
(\mu\,\sigma\,|\,\mathcal{B}) & \sim p(\mu\,\sigma\,|\,\mathcal{B}) \\
(\theta_i\,|\,\mu\,\sigma\,\mathcal{N}\,\mathcal{B}) & \overset{\text{IID}}{\sim} N(\mu,\sigma^2) \\
(y_i\,|\,\theta_i\,\mu\,\sigma\,V_i\,\mathcal{B}) & \overset{\text{I}}{\sim} N(\theta_i, V_i)
\end{aligned} \tag{15}
$$

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);
- 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 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 them 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 (15). This assumption, denoted by $\mathcal{N}$ 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; even so, it leads to useful results, as we will 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) 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 $\boldsymbol{y} = (y_1,\dots,y_k)$ and $\boldsymbol{V} = (V_1,\dots,V_k)$. It can be shown (You're not asked to show this; the calculation is made by (in the jargon) *integrating out the random effects θ_i*) that the likelihood function for $\boldsymbol{\eta} \triangleq (\mu,\sigma)$ in model (15) is

$$
\ell(\mu\,\sigma\,|\,\boldsymbol{y}\,\boldsymbol{V}\,\mathcal{N}\,\mathcal{B}) = \prod_{i=1}^{k} \frac{1}{\sqrt{V_i+\sigma^2}} \exp\left[-\frac{1}{2}\,\frac{(y_i-\mu)^2}{V_i+\sigma^2}\right] , \tag{16}
$$

leading to the log-likelihood function

$$
\ell\ell(\mu\,\sigma\,|\,\boldsymbol{y}\,\boldsymbol{V}\,\mathcal{N}\,\mathcal{B}) = -\frac{1}{2}\sum_{i=1}^{k}\left[\log(V_i+\sigma^2) + \frac{(y_i-\mu)^2}{V_i+\sigma^2}\right] . \tag{17}
$$

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

$$(\hat{\boldsymbol{\eta}} \,|\, D \, \mathcal{B}) \sim N_k\left(\boldsymbol{\eta}, \hat{I}^{-1}\right) , \tag{18}$$

in which the Fisher 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 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) 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 on the course web page, using the (μ, σ) grid mentioned above. 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: this meta-analysis has a total of $N = 10{,}816$ patients but only $k = 6$ studies. It turns out that the effective sample sizes for μ and σ are driven mainly by N and k, respectively. Do Your plots resemble the large-sample bivariate Normal shapes described above? Explain briefly. *[5 points]*

  (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]*

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 $k > 1$ and the problem is regular (in the exponential-family sense), one standard approach to obtain the MLEs, applied to the aspirin meta-analysis, involves (a) creating a system of 2 equations in 2 unknowns by setting each of the first partials with respect to μ and σ equal to 0 and (b) solving for (μ, σ). Sometimes these equations will have closed-form algebraic solutions, but more often in two or more dimensions they have to be solved numerically.

- The log-likelihood here is a function $\ell\ell\colon \mathbb{R}^k \to \mathbb{R}$ that takes as input a vector $\boldsymbol{\eta}$ of real numbers of length k 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 on the course web page; let's look at how this works, starting with `optim` first.

(d) *[45 total points for this sub-problem]* 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]*

(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. 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

$$\hat{\mu}_{MLE} - \Phi^{-1}(1-\alpha) \cdot \widehat{SE}\,(\hat{\mu}_{MLE}) \tag{19}$$

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 (19). Compute this LCB for $\alpha = 0.05$. *[5 points]* At the 95% 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]*

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 (15) is

$$\hat{\mu} = \frac{\sum_{i=1}^k \hat{W}_i\, y_i}{\sum_{i=1}^k \hat{W}_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^k \hat{W}_i^2\, [(y_i - \hat{\mu})^2 - V_i]}{\sum_{i=1}^k \hat{W}_i^2}\,, \quad \text{in which} \quad \hat{W}_i = \frac{1}{V_i + \hat{\sigma}^2}\,. \tag{20}$$

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 just set $\hat{\sigma}^2_{MLE} = 0$). 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 $\hat{\mu}_{MLE}$:

$$\widehat{SE}\,(\hat{\mu}_{MLE}) = \left[\sum_{i=1}^k \frac{1}{V_i + \hat{\sigma}^2_{MLE}} \right]^{-\frac{1}{2}} \tag{21}$$

(e) *[10 total points for this sub-problem]* I've written R code to implement this algorithm and posted it on the course web page. 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]*

(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]*

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

$$(\theta_i \,|\, y_i \, \mu \, \sigma \, \mathcal{N} \, \mathcal{B}) \overset{\text{I}}{\sim} N\left[\theta_i^*, V_i(1 - B_i)\right] \,, \quad \text{with} \quad \theta_i^* = (1 - B_i)\, y_i + B_i \, \mu \quad \text{and} \quad B_i = \frac{V_i}{V_i + \sigma^2} \,. \tag{22}$$

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 (15) 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

$$\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2} \quad \text{and} \quad \hat{\theta}_i = (1 - \hat{B}_i)\, y_i + \hat{B}_i \, \hat{\mu} \,, \tag{23}$$

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

$$\widehat{SE}\left(\hat{\theta}_i\right) = \sqrt{V_i \, (1 - \hat{B}_i)} \,. \tag{24}$$

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			0.990
2		0.141		0.202	0.529	2.50	1.94	0.899
3	626	0.0579		0.0934	0.782	1.84	1.53	
4	1682		0.232		0.646	2.56	1.84	0.994
5		0.112	0.183	0.120	0.719			1.04
6	4524		0.427	0.280	0.346	−1.15	−0.251	

(f) *[30 total points for this sub-problem]* Use the output from Your previous code execution 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 (23) 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]*

(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]*

(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]*

In the rest of this problem You'll perform a Bayesian analysis of the data in Table 2. Looking back at equation (15), 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 I want to build a low-information-content 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), so I'm going to 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

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	
		Standard Error		Posterior	
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		
θ_2		—	0.8995	2.042	
θ_3	1.533	—		1.592	1.542
θ_4	1.841	—	0.9941		1.315
θ_5		—	1.049	1.812	1.431
θ_6	-0.2514	—	0.7278	-0.4327	0.9425

$(\mu \,|\, \mathcal{B}) \sim \text{Uniform}(A, B)$, 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 course web page; after some experimentation I chose $(A, B, C) = (-2, 5, 6)$ in the prior specification. Run my code (or an equivalent program in some other language) 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.

(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 (15), has shown that Bayes and ML findings will either be similar (when k is large) or the ML approach will often underestimate uncertainty when it differs from Bayes. Does the second of those two possibilities appear to have happened here? Explain briefly. *[5 points]*

(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 95% Bayesian posterior interval for σ from the output and report it here. *[5 points]* Compute the large-sample-approximate 95% 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 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 them? Does this mean that σ is statistically significantly different from 0? Explain briefly. *[10 points]*

(iii) Show (by extracting the relevant number from Your output) that, conditional on model (15) 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]*