Prof. David Draper
Department of
Applied Mathematics and Statistics
University of California, Santa Cruz

AMS 132: Take-Home Final Exam (final draft)

Absolute due date: 24 Mar 2017 *[640 total points]*

As with Homeworks 1 and 2, please collect {all of the `R` 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, at the beginning of the Appendix you can say something like the following:

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

1. *[50 points]* For each statement below (10 points each), say whether it's true or false; if true without further assumptions, briefly explain why its true (and — extra credit (5 points each time) — 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.

In answering these questions you may find it helpful to consult the following references, available on the course web page: DS (Degroot and Schervish (2012)) sections 3.10, 12.5, 12.6; Gelman et al. (2014) Chapter 11.

(a) If you can figure out how to do IID sampling from the posterior distribution of interest to you, this will often be more Monte-Carlo efficient than MCMC sampling from the same posterior.

(b) A (first-order) Markov chain is a particularly simple stochastic process: to simulate where the chain goes next, you only need to know (i) where it is now and (ii) where it was one iteration ago.

(c) The bootstrap is a frequentist simulation-based computational method that can be used to create approximate confidence intervals for population summaries even when the population distribution of the outcome variable y of interest is not known; for example, if all you know from problem context is that your observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ are IID from *some* distribution with finite mean μ and finite SD σ, you can use the bootstrap to build an approximate confidence interval for μ even though you don't know what the population distribution is.

(d) Simulation-based computational methods are needed in Bayesian inference because conjugate priors don't always exist and high-dimensional probability distributions are difficult to summarize algebraically.

(e) In MCMC sampling from a distribution, you have to be really careful to use a burn-in period of just the right length, because if the burn-in goes on for too long the Markov chain will have missed its chance to find the equilibrium distribution.

2. *[210 points]* 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 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 that 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 (see DS section 5.9). Suppose that a population of interest contains items of $p \geq 2$ types (in the example here: people who support {Bush, Dukakis, other}, so that in this case $p = 3$) and that the population proportion of items of type j is $0 < \theta_j < 1$. Letting $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, note that there's a restriction on the components of $\boldsymbol{\theta}$, namely $\sum_{j=1}^{p} \theta_j = 1$. 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 p (n_p); let $\boldsymbol{N} = (n_1, \ldots, n_p)$ 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 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

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

The main scientific 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) 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

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

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

(b) Returning now to the general Multinomial setting, briefly explain why the likelihood function for $\boldsymbol{\theta}$ given $\boldsymbol{N}$ is

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

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

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

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

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

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 $\sum_{j=1}^{p} \theta_j = 1$. There are thus two ways forward to compute the MLE:

(i) Solve the constrained optimization problem directly with *Lagrange multipliers (Extra credit [20 points]: do this)*, or

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

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

from which (ignoring the irrelevant constant)

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

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

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

The MLE for $(\theta_1, \ldots, \theta_{p-1})$ may now be found by setting $\frac{\partial}{\partial\, \theta_j} \, \ell\ell(\theta_1, \ldots, \theta_{p-1} \,|\, \boldsymbol{N}\, \mathcal{B}) = 0$ for $j = 1, \ldots, (p-1)$ and solving the resulting system of $(p-1)$ equations in $(p-1)$ unknowns *(Extra credit [20 points]: do this for general p)*, but that gets quite messy; let's just do it for $p = 3$, 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

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

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

(c) It can be shown *(Extra credit [20 points]: do this for general p, 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 the estimated covariance matrix of the MLE vector $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right)$ is

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

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 $p = 2$ outcome categories); also explain why it makes good sense that the off-diagonal elements of $\hat{\boldsymbol{\Sigma}}$ are negative. 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

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

Finally, use your estimated SE for $\hat{\gamma}$ to construct an approximate (large-sample) 95% confidence interval for γ. Was Bush definitively ahead of Dukakis at the point when the survey was conducted? Explain briefly. *[50 points]*

(d) 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_p to a (possibly different) power.

There is such a distribution — it's called the *Dirichlet*($\boldsymbol{\alpha}$) distribution, with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ chosen so that all of the α_i are positive:

$$p(\boldsymbol{\theta}\,|\,\mathcal{B}) = c\,\prod_{i=1}^{p} \theta_i^{\alpha_i - 1}\,. \tag{12}$$

Briefly explain why this means that the conjugate updating rule is

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

Given that $\boldsymbol{N} = (n_1, \ldots, n_p)$ and that the n_j represent sample sizes (numbers of observations y_i) in each of the p Multinomial categories, briefly explain why this implies that, if context suggests a low-information-content prior, this would correspond to choosing the α_i all close to 0. *[20 points]*

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

(i) pick any $\beta > 0$ of your choosing ($\beta = 1$ is a good choice that leads to fast random number generation);

(ii) for $j = 1, \ldots, p$ make p independent draws g_j with draw j from the $\Gamma(\alpha_j, \beta)$ distribution; and

(iii) then just normalize:

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

Write Your own code, using this algorithm, to generate M IID draws from the posterior distribution implied 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 $\gamma = (\theta_1 - \theta_2)$. 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 small enough to make all of the posterior mean estimates reliable to at least 3 significant figures, and justify your choice. 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. How do Your Bayesian answers compare with those from maximum likelihood in this problem? Explain briefly. Compute a Monte Carlo estimate of $p(\gamma > 0 \,|\, \boldsymbol{N}\,\mathcal{B})$, 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. 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. *[100 points] Extra credit ([10 points]): Use* `Maple` *or some equivalent environment (or paper and pen, if You're brave) to see if You can derive a closed-form expression for* $p(\gamma > 0 \,|\, \boldsymbol{N}\,\mathcal{B})$*, and compare Your mathematical result with Your simulation-based findings; if no such expression is possible, explain why not.*

3. *[380 points]* The data set `U.S.-family-income-2009.txt` on the course web site contains family income values $\boldsymbol{y} = (y_1, \ldots, y_n)$ from the year 2009 (in units of \$1,000, and sorted from smallest to largest) for a random sample of $n = 842$ American households with incomes under \$1 million, obtained from the Current Population Survey conducted by the *U.S. Census Bureau*. The point of this problem is to draw inferences about the mean family income θ in the population $\mathcal{P}$ of all U.S. households (with incomes under \$1 million) in 2009. In estimating θ we'll look at two frequentist-and-Bayesian parametric methods and two frequentist nonparametric approaches; the meta-goal of this problem is to experience statistical model uncertainty and to begin to learn how to cope with it.

(a) Make a histogram of these income values (on the density scale) with a large number of bars (e.g., 100); briefly comment on its shape, and visually estimate the mode. Compute the sample mean, median and SD, and compare the mean, median and mode; are they related to each other in the way you would expect them to be, given the distributional shape? Explain briefly. *[40 points]*

Two common parametric sampling models for long-right-hand-tailed variables such as family income are based on the Lognormal and Gamma distributions. Consider first the model

$$\left\{ \begin{array}{c} (Y_i \,|\, \mu\, \sigma^2\, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Lognormal}(\mu, \sigma^2) \\ (i = 1, \ldots, n) \end{array} \right\}, \tag{15}$$

in which $Y_i \sim \text{Lognormal}(\mu, \sigma^2)$ simply means that $W_i \triangleq \log Y_i \sim N(\mu, \sigma^2)$. (You can see that Lognormal is a bad name for this distribution: if $W_i = \log Y_i$ is Normal then $Y_i = e^{W_i}$, so it should really be called the Exponential Normal model, but this name has been in use for more than 100 years and we're stuck with it.)

(b) It can be shown *(Extra credit [20 points]: show these two facts)* that in this model

(i) the repeated-sampling density of Y_i is

$$p(y_i \,|\, \mu\, \sigma^2 \mathcal{B}) = \frac{1}{\sigma\, y_i \sqrt{2\,\pi}} \exp\left[-\frac{(\log y_i - \mu)^2}{2\,\sigma^2} \right] \tag{16}$$

for $y_i > 0$ (and 0 otherwise) and

(ii) $E_{RS}(Y_i \,|\, \mu\, \sigma^2\, \mathcal{B}) = \exp\left(\mu + \frac{\sigma^2}{2}\right)$, meaning that if we adopt sampling model (15) then $\theta = \exp\left(\mu + \frac{\sigma^2}{2}\right)$.

The definition of the Lognormal distribution — $W_i \triangleq \log Y_i \sim N(\mu, \sigma^2)$ — and fact (i) suggest two different ways to obtain the MLEs $(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$ for μ and σ^2 in model (15). Describe both of these ways to get the ML estimators; identify which method is easier, and briefly explain why; and use the easier method to get the MLEs. Plot your histogram of the data set on the density scale from part (a) again, and superimpose a density trace of the Lognormal$(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$ distribution; does this fit look acceptable? Explain briefly (*Hint:* You can either write your own Lognormal density function (from equation (16)) or gain access to the Lognormal density function `dlnorm` in `R` (and the information about it available with the `help` command) by first issuing the command `library( stats )`.) *[60 points]*

(c) Use `RJAGS` to fit, via MCMC, a Bayesian version of the Lognormal model to the income data with a low-information-content prior. *I'll provide all of the code you need to do this on the course web page.* It turns out that to get DIC values for Bayesian models in `RJAGS`, you need to run $k = 2$ or more parallel Markov chains with different random number streams, so when you ask for M iterations in `RJAGS` with $k = 2$ you're actually getting $2\,M$ iterations, so I recommend a burn-in of 1,000 iterations from the starting values I've given you, followed by a monitoring run of $M = 50{,}000$ iterations (on my desktop at home this only took about 1.5 minutes). Compare the posterior mean values for μ and σ^2 with their corresponding MLEs, and comment briefly on whether the comparison came out as you expected it to. Summarize the posterior for θ by extracting its posterior mean, SD and 95% interval from the MCMC output. Compare the mean and SD and several quantiles of the predictive distribution for a new y value with the corresponding quantities from the income data set, and make a qqplot of your random draws from the predictive distribution against the data set; does the model seem to make good predictions? Explain briefly. Get `RJAGS` to work out the DIC value for the Lognormal model. *[60 points]*

Now instead consider the model (for $\alpha > 0, \beta > 0$)

$$\left\{ \begin{array}{c} (Y_i \,|\, \alpha\, \beta\, \mathcal{B}) \overset{\text{IID}}{\sim} \Gamma(\alpha, \beta) \\ (i = 1, \ldots, n) \end{array} \right\} \quad \longleftrightarrow \quad p(y_i \,|\, \alpha\, \beta\, \mathcal{B}) = \left\{ \begin{array}{ll} \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta\, y_i} & \text{for } y_i > 0 \\ 0 & \text{otherwise} \end{array} \right\}, \tag{17}$$

using the parameterization in which $E_{RS}(Y_i \,|\, \alpha\, \beta\, \mathcal{B}) = \frac{\alpha}{\beta}$, meaning that if we adopt sampling model (17) then $\theta = \frac{\alpha}{\beta}$.

(d) Show that the log-likelihood function in this model is

$$\ell\ell(\alpha\, \beta \,|\, \boldsymbol{y}\, \mathcal{B}) = n\, \alpha\, \log \beta - n\, \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log y_i - \beta \sum_{i=1}^{n} y_i \,. \tag{18}$$

Briefly explain why this means that no closed-form (algebraic) expressions for the MLEs for α and β are possible in this model; numerical methods are needed to maximize (18). *(Extra credit ([20 points]): Figure out how to use the* `R` *function* `optim` *to numerically maximize the log-likelihood function; note that* `optim` *has a* `Hessian = TRUE` *option that will deliver to you not only the MLEs but also the Hessian evaluated at the MLEs.)* It turns out with this data set that $(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE}) \doteq (1.305, 0.01575)$. Plot your histogram of the data set on the density scale from part (a) yet again, and superimpose a density trace of the $\Gamma(\hat{\alpha}_{MLE}, \hat{\beta}_{MLE})$ distribution; does *this* fit look acceptable? Explain briefly. *[30 points]*

(e) Now use `RJAGS` to fit, via MCMC, a Bayesian version of the Gamma model to the income data with a low-information-content prior. *Again, I'll provide all of the code you need to do this on the course web page.* This model takes longer to fit (this fact is related to the difficulty you saw in part (d) with maximum-likelihood fitting), so you may wish to settle for a smaller value of M (e.g., 10,000) than you did with the Lognormal model (on my desktop at home, 20,000 monitoring iterations took about 70 seconds; you can adjust your value of M based on the speed of your machine and your patience). As you did in the Lognormal case, summarize the posterior for θ by extracting its posterior mean, SD and 95% interval from the MCMC output. Compare the mean and SD and several quantiles of the predictive distribution for a new y value with the corresponding quantities from the income data set, and make a qqplot of your random draws from the predictive distribution against the data set; does *this* model seem to make good predictions? Explain briefly. Get `RJAGS` to work out the DIC value for the Lognormal model. *[60 points]*

(f) Since interest focuses on the population mean θ, one possible estimator of θ is the sample mean $\bar{Y}$. Use the Central Limit Theorem (CLT) to construct an approximate 95% confidence interval for θ based on $\bar{Y}$. *[10 points]*

Notice that the interval you created in part (f) is based on a frequentist nonparametric method: in building your interval you made no assumptions about the form of the sampling distribution. **Question:** how well calibrated is the CLT interval? By this I mean the usual notion of validity/calibration of 95% intervals: if the process leading to your interval in part (f) is repeated many times, do the resulting intervals include the true parameter value θ approximately 95% of the time?

In answering this question it looks like we need to know what the truth is about θ. Since we don't know θ, it turns out that the best we can do – in estimating the calibration of the CLT interval

Table 2: *Summary of inference about θ across six models/methods; ML = maximum likelihood.*

	θ		
Method/Model	Mean/ Estimate	SD/ SE	95% Interval
Lognormal ML	·	—	—
Lognormal Bayes	·	·	$(\cdot, \cdot)$
Gamma ML	·	—	—
Gamma Bayes	·	·	$(\cdot, \cdot)$
CLT	·	·	$(\cdot, \cdot)$
Bootstrap	·	·	$(\cdot, \cdot)$

process – is to use the bootstrap idea we talked about in class: we can take random samples from our *sample* and use these as proxies for what we would get if we were able to take a new random sample of $n = 842$ households from the population $\mathcal{P}$. In this bootstrapping, we'll pretend that the true θ is the sample mean $\bar{y} \doteq 82.88$ you calculated in part (a).

(g) Write a program in `R` to repeatedly ($M = 10{,}000$ times, say) draw samples of size $n = 842$ at random with replacement from the data set `U.S.-family-income-2009.txt`, compute the CLT-based 95% interval each time, and work out the proportion of these intervals that include the actual sample mean (put a copy of your code in the Appendix, as usual). To get ready for parts (h) and (i), also keep track of the M simulated sample means $\bar{y}^*$ you calculated in assessing calibration of the CLT interval-building process. *[30 points]*

(h) The CLT method assumes that $n = 842$ is large enough to produce a Normal sampling distribution for $\bar{Y}$. Check this assumption by making a normal qqplot of the $\bar{y}^*$ values you generated in part (g) — has the CLT done its magic with $n = 842$, even though the population distribution had a quite long right-hand tail? Explain briefly. *[20 points]*

(i) In producing the bootstrap distribution of the sample means $\bar{y}^*$ in part (g), you've also done all of the hard work to construct the bootstrap confidence interval for θ, which you can now obtain simply by getting `R` to work out the 2.5% and 97.5% quantiles of the $\bar{y}^*$ distribution. How does this bootstrap CI compare with the CLT-based CI? Explain briefly. *[20 points]*

(j) Summarize all of this by completing Table 2; a dash (—) in an entry means that you don't need to fill it in. Given your conclusions about model fit and validity of method, which numbers in Table 2 seem most reliable to you? What do you conclude, both about mean income in the U.S. in 2009 and about the challenge of model uncertainty? *[50 points]*