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

AMS 132: Homework 2 (final version)

Target due date: Tue 14 Mar 2017 *[420 total points]*

As with Homework 1, 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 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. *[60 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.

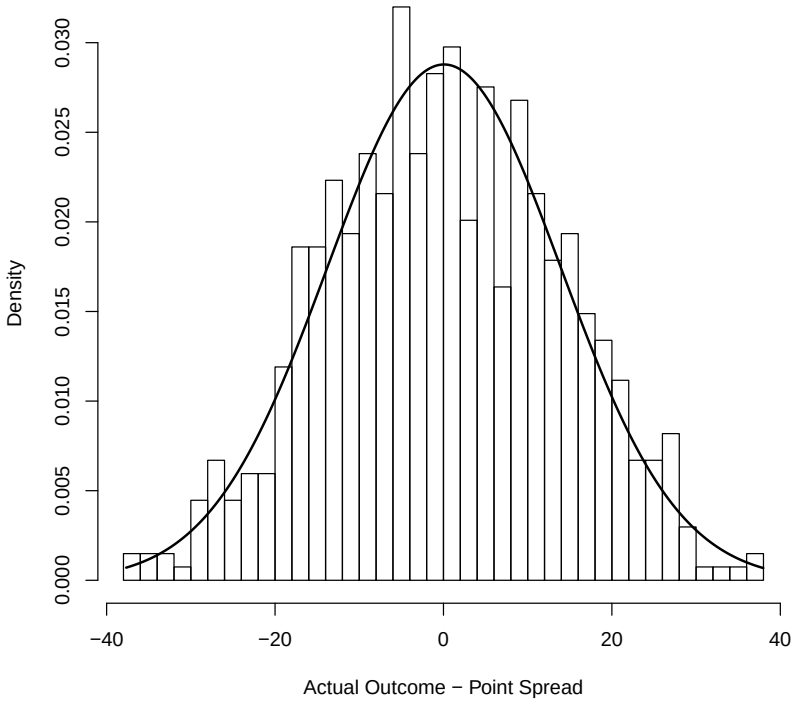
(a) Consider the sampling model $(Y_i \,|\, \theta\, \mathcal{B}) \stackrel{\text{IID}}{\sim} p(y_i \,|\, \theta\, \mathcal{B})$ for $i = 1, \ldots, n$, where the Y_i are univariate real values, θ is a parameter vector of length $1 \leq p < \infty$ and $\mathcal{B}$ summarizes Your background information; a Bayesian analysis with the same sampling model would add a prior distribution layer of the form $(\theta \,|\, \mathcal{B}) \sim p(\theta \,|\, \mathcal{B})$ to the hierarchy. The Bernstein-von Mises Theorem says that maximum-likelihood (ML) and Bayesian inferential conclusions will be similar in this setting if (a) n is large and (b) $p(\theta|\mathcal{B})$ is diffuse, 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) Being able to express Your sampling distribution as a member of the exponential family is helpful, because (1) You can then readily identify a set of sufficient statistics just by looking at the form of the exponential family and (2) the conjugate prior is also directly available from the exponential family form.

(c) 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 6 of the extra notes from 17 Jan 2017)), 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.

(d) When Your sampling model has n observations and a single parameter θ (so that $p = 1$), 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 (1) information in $\hat{\theta}_{MLE}$ about θ increases linearly with n and (2) $\hat{V}\left(\hat{\theta}_{MLE}\right) = O\left(\frac{1}{n}\right)$.

Figure 1: *Differences y_i between observed and predicted American football scores, 1981–1984.*

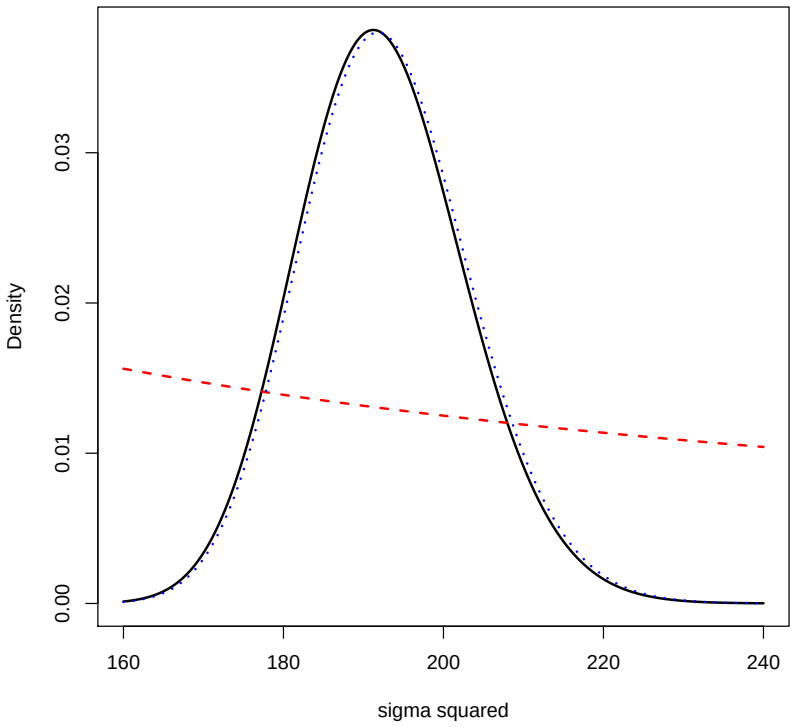


(e) It's easier to reason from the part (or the particular, or the sample) to the whole (or the general, or the population), and that's why statistical inference (inductive reasoning) is easier than probability (deductive reasoning).

(f) When the sampling model is a regular parametric family $p(Y \,|\, \theta\, \mathcal{B})$, where θ is a vector of length $1 \leq p < \infty$ and $Y = (Y_1, \ldots, Y_n)$, for large n the repeated-sampling distribution of the (vector) MLE $\hat{\theta}_{MLE}$ is approximately p–variate normal with mean vector θ and covariance matrix $\hat{I}^{-1}$ (the inverse of the observed information matrix), and the bias of $\hat{\theta}_{MLE}$ as an estimate of θ in large samples is $O\big(\frac{1}{n^2}\big)$.

2. *[200 points]* People in Las Vegas who are experts on the National Football League provide a *point spread* for every football game before it occurs, as a measure of the difference in ability between the two teams (and taking account of where the game will be played). For example, if Denver is a 3.5–point favorite to defeat San Francisco, the implication is that betting on whether Denver's final score minus 3.5 points exceeds or falls short of San Francisco's final score is an even-money proposition. With the definition *actual outcome = (score of favorite – score of underdog)*, Figure 1 (based on data from Gelman et al. (2014)) presents a histogram of the differences $y =$ (actual outcome – point spread) for a sample of $n = 672$ professional football games in the early 1980s, with a normal density superimposed having the same mean $\bar{y} = 0.07$ and standard deviation (SD) $s_u = 13.86$ as the sample (if this distribution didn't have a mean that's close to 0, the experts would be *uncalibrated* and you could make money by betting against them). You can see from this figure that the model $(Y_i \,|\, \sigma^2\, \mathcal{B}) \overset{\text{IID}}{\sim} N(0, \sigma^2)$ is reasonable for the observed differences y_i.

Figure 2: *Prior, likelihood, and posterior densities with the football data of Figure 1.*



(a) Write down the likelihood and log likelihood functions for σ^2 in this model. Show that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$, which takes the value 191.8 with the data in Figure 1, is both sufficient and the maximum likelihood estimator (MLE) for σ^2. Plot the log likelihood function for σ^2 in the range from 160 to 240 with these data, briefly explaining why it should be slightly skewed to the right. *[60 points]*

(b) Show that the conjugate prior for σ^2 in this model is the *scaled inverse chi-square* distribution,

$$(\sigma^2|\mathcal{B}) \sim \chi^{-2}(\nu_0, \sigma_0^2), \quad \text{i.e.,} \quad p(\sigma^2|\mathcal{B}) = c\,\left(\sigma^2\right)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\nu_0\,\sigma_0^2}{2\,\sigma^2}\right), \tag{1}$$

where ν_0 is the prior sample size and σ_0^2 is a prior estimate of σ^2. In an attempt to be "non-informative" people sometimes work with a version of (1) obtained by letting $\nu_0 \to 0$, namely $p(\sigma^2|\mathcal{B}) = c_0\,(\sigma^2)^{-1}$. The resulting prior is *improper* in that it integrates to ∞, but it turns out that posterior inferences will be sensible nonetheless (even with sample sizes as small as $n = 1$). Show that with this prior, the posterior distribution is $\chi^{-2}(n, \hat{\sigma}^2)$. Given the interpretation of ν_0 as the prior sample size and σ_0^2 as the prior estimate of σ^2, do the values of n (for ν) and $\hat{\sigma}^2$ (for σ^2) in the posterior $\chi^{-2}(n, \hat{\sigma}^2)$ make good intuitive sense? Explain briefly. *[30 points]*

(c) Figure 2 plots the prior, likelihood, and posterior densities on the same graph using the data in Figure 1 and taking $c_0 = 2.5$ for convenience in the plot. Get `R` to reproduce this figure (**NB** `Maple` has a hard time doing this). You'll need to be careful to use the correct normalizing constant c in (1), which can be found in the lecture notes and in Appendix A of Gelman et

Table 1: *Comparison of frequentist and Bayesian inference about σ^2 with the football data of Figure 1.*

Frequentist			Diffuse-Prior Bayesian	
Estimate	Large-Sample 95% Interval	Small-Sample 95% Interval	Estimate	95% Interval
·	$(\cdot, \cdot)$	$(\cdot, \cdot)$	·	$(\cdot, \cdot)$

al. (2014); and because the data values in this example lead to astoundingly large and small numbers on the original scale, it's necessary to do all possible computations on the log scale and wait to transform back to the original scale until the last possible moment (you'll need to use the built-in function `lgamma` in `R`). Explicitly identify the three curves, and briefly discuss what this plot implies about the updating of information from prior to posterior in this case. *[40 points]*

(d) Fill in the dots in Table 1 by making the following computations.

- You already know the frequentist estimate, from part (i); to get the Bayesian estimate, let's use the posterior mean, which you can work out from the fact (mentioned in class) that

$$\text{if} \quad (\sigma^2 \,|\, \mathcal{B}) \sim \chi^{-2}(\nu_0, \sigma_0^2) \quad \text{then} \quad E(\sigma^2 \,|\, \mathcal{B}) = \left(\frac{\nu_0}{\nu_0 - 2} \right) \sigma_0^2 \quad \text{as long as} \quad \nu_0 > 2 \,. \tag{2}$$

- Work out the Fisher information from your log-likelihood function in (i) and use it to construct the large-sample frequentist 95% confidence interval for σ^2.
- In a small modification of Mr. Gosset's result (from class) about the small-sample-exact confidence interval for σ^2 in the Gaussian sampling model when μ is unknown, it can also be shown (you're not asked to show this) that under the model $(Y_i \,|\, \sigma^2 \,\mathcal{B}) \overset{\text{IID}}{\sim} N(0, \sigma^2)$ for $i = 1, \ldots, n$ (which we're using for the football data), the exact $100(1-\alpha)\%$ confidence interval for σ^2 is of the form

$$\left[\frac{n\,\hat{\sigma}_u^2}{(\chi^2_n)_{1-\frac{\alpha}{2}}}, \frac{n\,\hat{\sigma}_u^2}{(\chi^2_n)_{\frac{\alpha}{2}}} \right] , \tag{3}$$

  in which $\hat{\sigma}_u^2 = \frac{1}{n} \sum_{i=1}^n y_i^2$ is an unbiased estimate of σ^2 (here this coincides with the MLE) and $(\chi^2_\nu)_\gamma$ is the place along the χ^2 curve with ν degrees of freedom where $100\,\gamma\%$ of the total area under the curve is to the left of that place (i.e., the γ quantile of the χ^2_ν distribution).
- To get the 95% Bayesian interval, use the fact, noted in class, that the $\chi^{-2}(\nu_0, \sigma_0^2)$ density is the same as the $\Gamma^{-1}\left(\frac{\nu_0}{2}, \frac{\nu_0\,\sigma_0^2}{2}\right)$ distribution, and make Inverse Gamma calculations with the `qinvgamma` function in the `CRAN` package `actuar` (see the item called *R code for the Bayesian Gaussian analysis of the NB10 data* on the course web page for an example of this).

Briefly summarize how the frequentist and Bayesian results are similar and how they differ. Is this an example of the Bernstein-von Mises Theorem in action? Expain briefly. *[70 points]*

3. *[160 points]* Paleobotanists estimate the moment in the remote past when a given species became extinct by taking cylindrical, vertical core samples well below the earth's surface and looking for the last occurrence of the species in the fossil record, measured in meters above the point P at which the species was known to have first emerged. Letting $\boldsymbol{y} = (y_1, \ldots, y_n)$ denote a sample of such distances above P at a random set of locations, the sampling model $\boxed{(Y_i \,|\, \theta\, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Uniform}(0, \theta)\ (*)}$ emerges from simple and plausible assumptions. In this model the unknown $\theta > 0$ can be used, through carbon dating, to estimate the species extinction time. This problem is about likelihood and Bayesian inference for θ in model $(*)$, and it will be seen that some of our usual intuitions (derived from the Bernoulli, Poisson, and Gaussian case studies) do not hold in this case.

The marginal sampling distribution of a single observation y_i in this model may be written

$$p(y_i \,|\, \theta\, \mathcal{B}) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{if } 0 \le y_i \le \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta}\, I(0 \le y_i \le \theta)\,, \tag{4}$$

where (as usual) $I(A) = 1$ if proposition A is true and 0 otherwise.

(a) Use the fact that ($0 \le y_i \le \theta$ for all $i = 1, \ldots n$) if and only if ($m \triangleq \max(y_1, \ldots y_n) \le \theta$) to show that the likelihood function in this model is

$$\ell(\theta \,|\, \boldsymbol{y}\, \mathcal{B}) = \theta^{-n}\, I(\theta \ge m)\,. \tag{5}$$

Briefly explain why this demonstrates that m is sufficient for θ in this model. *[20 points]*

(b) As we've discussed in class, the maximum likelihood estimator (MLE) of a parameter θ is the value of θ (which will be a function of the data) that maximizes the likelihood function, and this maximization is usually performed by setting the derivative of the likelihood (or log likelihood) function to 0 and solving. Show by means of a rough sketch of the likelihood function in (a) that m is the maximum likelihood estimator (MLE) of θ, and briefly explain why the usual method for finding the MLE fails in this case. *[20 points]*

(c) A positive quantity θ follows the *Pareto* distribution with parameters $\alpha, \beta > 0$ — written $(\theta \,|\, \mathcal{B}) \sim \text{Pareto}(\alpha, \beta)$, and named for the Italian economist Vilfredo Pareto (1848–1923) — if it has density

$$p(\theta \,|\, \mathcal{B}) = \left\{ \begin{array}{cc} \alpha\, \beta^{\alpha}\, \theta^{-(\alpha+1)} & \text{if } \theta \ge \beta \\ 0 & \text{otherwise} \end{array} \right\}\,. \tag{6}$$

This distribution has mean $\frac{\alpha\,\beta}{\alpha-1}$ (if $\alpha > 1$) and variance $\frac{\alpha\,\beta^2}{(\alpha-1)^2\,(\alpha-2)}$ (if $\alpha > 2$). With the likelihood function viewed as (a constant multiple of) a density for θ, show that the likelihood corresponds to the Pareto$(n-1, m)$ distribution. Show further that if the prior distribution for θ is taken to be (6), under the model $(*)$ above the posterior distribution is $p(\theta \,|\, \boldsymbol{y}\, \mathcal{B}) = \text{Pareto}[\alpha + n, max(\beta, m)]$, thereby demonstrating that the Pareto distribution is conjugate to the Uniform$(0, \theta)$ likelihood. *[20 points]*

(d) In an experiment conducted in the Antarctic in the 1980s to study a particular species of fossil ammonite, the following was a linearly rescaled version of the observed data: $\boldsymbol{y} = (2.8, 1.7, 1.0, 5.1, 3.7, 1.5, 4.3, 2.0, 3.2, 2.1, 0.4)$. Prior information equivalent to a Pareto prior specified by the choice $(\alpha, \beta) = (2.5, 4)$ was available. Plot the prior, likelihood, and posterior distributions arising from this data set on the same graph, explicitly identifying the three curves, and briefly discuss what this picture implies about the updating of information from prior to posterior in this case. *[30 points]*

(e) Make a table summarizing the mean and standard deviation (SD) for the prior (Pareto(α, β)), likelihood (Pareto$(n-1, m)$), and posterior (Pareto$[\alpha+n, \max(\beta, m)]$) distributions, using the (α, β) choices and the data in part (d) above. In Bayesian updating the posterior mean is usually (at least approximately) a weighted average of the prior and likelihood means (with weights between 0 and 1), and the posterior SD is typically smaller than either the prior or likelihood SDs. Are each of these behaviors true in this case? Explain briefly. *[50 points]*

(f) You've shown in part (c) that the posterior for θ based on a sample of size n in model $(*)$ is $p(\theta \,|\, \boldsymbol{y}\, \mathcal{B}) = \text{Pareto}[\alpha+n, \max(\beta, m)]$. Write down a symbolic expression for the posterior variance of θ in terms of (α, β, m, n). When considered as a function of n, what's unusual about this expression in relation to the findings in our previous case studies in this course? Explain briefly. *[20 points]*