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

AMS 132: Homework 1

Target due date: Thu 16 Feb 2017 *[330 total points]*

Here's a style guide for all of the written work in this class. In figuring out how to write up answers to homework and final exam problems, pretend the grader is sitting there with you and you're having a brief discussion with her/him on each question — that is, write down in a few sentences what you would say to someone to support your position. It's never enough in this class to just say "yes" or "10.3," even if the right answer is "yes" or "10.3"; you need to say "yes (or 10.3), because ... ." The right answer with no reasoning to support it, or the wrong reasoning, will get about half credit in this course, as will the wrong answer arrived at with a good effort. Leaving a problem or a part of a problem blank will get no credit.

1. *[110 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) You're about to spin a roulette wheel, which will result in a metal ball landing in one of 38 slots numbered $\Omega = \{0, 00, 1, 2, ..., 36\}$; 18 of the numbers from 1 to 36 are colored red, 18 are black, and 0 and 00 are green. You regard this wheel-spinning as fair, by which You mean that all 38 elemental outcomes in Ω are equipossible. Under Your assumption of fairness, the classical (Pascal–Fermat) probability of getting a red number on the next spin exists, is unique, and equals $\frac{18}{38}$.

(b) Under the same conditions as (a), the Kolmogorov (frequentist) probability of getting a red number on the next spin exists, is unique, and equals $\frac{18}{38}$.

(c) Under the same conditions as (a), the Bayesian probability of getting a red number on the next spin — with the fairness of the roulette wheel as part of your background assumptions in $\mathcal{B}$ — exists, is unique, and equals $\frac{18}{38}$.

(d) You're a professor; a new student (whom You've never met before) comes to Your office on the day before the quarter begins, saying that she (the student) wants to take a class that You're about to teach that quarter, but she's worried she may fail. The Kolmogorov (frequentist) probability that she will fail the class, if she takes it, is undefined, because there's no unique Ω on which to base the Kolmogorov probability calculation.

(e) In the Bernoulli sampling model, in which $(Y_i \,|\, \theta\, \mathcal{B}) \stackrel{\text{IID}}{\sim} \text{Bernoulli}(\theta)$ for $i = 1, \ldots, n$, the sum $s_n = \sum_{i=1}^n y_i$ of the observed data values $\boldsymbol{y} = (y_1, \ldots, y_n)$ is sufficient for inference about θ, and this means that in this model You can throw away the data vector $\boldsymbol{y}$ and focus only on s_n without any loss of information whatsoever.

(f) In learning how to do a good job on the task of uncertainty quantification, it's good to know quite a bit about both the Bayesian and frequentist paradigms, because (a) the Bayesian approach to probability ensures logical internal consistency of Your uncertainty assessments but does not guarantee good calibration, and (b) the frequentist approach to probability provides a natural framework in which to see if Your Bayesian answer *is* well-calibrated.

(g) The Beta$(\theta \,|\, \alpha\, \beta)$ parametric family of distributions is useful as a source of prior distributions when the sampling model is as in (d), because all distributional shapes (symmetric, skewed, multimodal, ...) on $(0, 1)$ are realizable in this family.

(h) Specifying the ingredients $\{p(\theta \,|\, \mathcal{B}), p(D \,|\, \theta\, \mathcal{B}), (\mathcal{A} \,|\, \mathcal{B}), U(a, \theta \,|\, \mathcal{B})\}$ in Your model for Your uncertainty about an unknown θ (in light of background information $\mathcal{B}$ and data D) is typically easy, because in any given problem there will typically be one and only one way to specify each of these ingredients; an example is the Bernoulli sampling distribution $p(D \,|\, \theta\, \mathcal{B})$ arising uniquely, under exchangeability, from de Finetti's Representation Theorem for binary outcomes.

(i) In trying to construct a good uncertainty assessment of the form $P(A \,|\, \mathcal{B})$, where A is a proposition and $\mathcal{B}$ is a set of propositions of the form (B_1 and B_2 and $\cdots$ and B_b), You should try hard not to condition on any propositions B_i that are false, because that would be the probabilistic equivalent of dividing by zero.

(j) The kind of objectivity in probability assessment sought by people like Venn, in which all reasonable people would agree on the assessed value, is often impossible to achieve, because all such assessments are conditional on the (1) assumptions, (2) information base and (3) judgments of the person making the probability assessment, and different reasonable people can differ along any of those three dimensions.

(k) One reason that Bayesian inference was not widely used in the early part of the 20th century was that approximating the (potentially high-dimensional) integrals arising from the Bayesian approach was difficult in an era when computing was slow and the Laplace-approximation technique had been forgotten.

2. *[220 points]* As a small part of a study I worked on at the RAND Corporation in the late 1980s, we obtained data on a random sample of $n = 14$ women who came to a hospital in Santa Monica, CA, in 1988 to give birth to premature babies. One outcome of interest was the length of stay (LoS) y_i in the hospital that woman i in this sample experienced, recorded as an integer; it was possible for this variable to be recorded as 0 if the LoS was under 12 hours. The data values were as follows: $\boldsymbol{y} = (y_1, \ldots, y_n) = (1, 2, 1, 1, 4, 1, 2, 2, 0, 3, 6, 2, 1, 3)$. The unknown θ of principal interest in this problem is the mean LoS for all women giving birth to premature babies at this Santa Monica hospital in 1988.

(a) Make a histogram or stem-and-leaf plot of the observed data values, and describe the basic shape of this distribution. *[10 points]*

One possible sampling model for non-negative integer-valued variables is the *Poisson* distribution with mean θ: You could take the $(Y_i \,|\, \theta\, \mathcal{B})$ as conditionally IID Poisson(θ), where the marginal

sampling distribution for observation i would then be

$$P(Y_i = y_i \,|\, \theta\, \mathcal{B}) = \left\{ \begin{array}{cc} \frac{\theta^{y_i}\, e^{-\theta}}{y_i!} & \text{for } y_i = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right\} \tag{1}$$

(here $\mathcal{B}$ would include the *assumption*, not automatically rendered true by the context of the problem, of the Poisson sampling distribution).

(b) Verify that, if the Poisson distribution is parameterized in this way, θ is indeed the mean of this distribution; in other words, show that if $(Y_i \,|\, \theta\, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Poisson}(\theta)$ then $E(Y_i \,|\, \theta\, \mathcal{B}) = \theta$. Use this to show that the method-of-moments estimator $\hat{\theta}_{MoM}$ of θ is the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. *[20 points]*

(c) Work out the likelihood and log likelihood functions in this model with this data set and plot them. How close are they in this small-sample case to the Gaussian behavior You'd expect (on Bayesian grounds) in large samples? Show that $S_n = \sum_{i=1}^{n} Y_i$ is sufficient for θ in this sampling model and that the maximum-likelihood estimator $\hat{\theta}_{MLE}$ of θ is $\bar{Y}$. *[50 points]*

(d) To get an informal idea of whether the Poisson sampling model fits this data set reasonably well, complete the following table:

	$\hat{P}(Y_i = y_i)$	
y_i	Empirical	Best-Fitting Poisson
0	0.071	
1		0.261
2		
3		
4		
5		
6		
≥ 7		

In this table, the empirical $\hat{P}(Y_i = y_i)$ values are just the observed relative frequen- cies, and the best-fitting Poisson values are obtained by computing probabilities from the Poisson distribution with $\theta = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Informally, does the fit look good to You? Explain briefly. *[30 points]*

(e) Verify that in this model θ is also the variance of the distribution, and use this to create another informal check on the Poisson sampling model. *[20 points]*

(f) Compute a large-sample standard error for $\hat{\theta}_{MLE}$ using observed Fisher information, and use this to construct an approximate 95% confidence interval for θ. *[20 points]*

(g) Show that the conjugate prior for θ in the Poisson sampling model is the Gamma distribution $\Gamma(\alpha, \beta)$: for $\alpha > 0$ and $\beta > 0$,

$$p(\theta \,|\, \alpha\, \beta\, \mathcal{B}) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\, \theta^{\alpha-1}\, e^{-\beta\, \theta}\,. \tag{2}$$

[10 points]

(h) Show that the conjugate updating rule in this model is

$$\left\{ \begin{array}{c} (\theta \,|\, \alpha\, \beta\, \mathcal{B}) \sim \Gamma(\alpha, \beta) \\ (Y_i \,|\, \theta\, \mathcal{B}) \overset{\text{IID}}{\sim} \text{Poisson}(\theta) \\ (i = 1, \ldots, n) \end{array} \right\} \rightarrow (\theta \,|\, \boldsymbol{y}\, \mathcal{B}) = (\theta \,|\, s_n\, \mathcal{B}) \sim \Gamma(\alpha + s_n, \beta + n)\,, \tag{3}$$

in which $s_n = \sum_{i=1}^{n} y_i$ is the observed value of the sufficient statistic S_n. *[10 points]*

(i) With this parameterization of the Gamma distribution, it turns out that

$$\text{if} \quad (\theta \,|\, \alpha\, \beta\, \mathcal{B}) \sim \Gamma(\alpha, \beta) \quad \text{then} \quad E(\theta \,|\, \alpha\, \beta\, \mathcal{B}) = \frac{\alpha}{\beta} \quad \text{and} \quad V(\theta \,|\, \alpha\, \beta\, \mathcal{B}) = \frac{\alpha}{\beta^2}\,. \tag{4}$$

Use the mean expression in equation (**??**) to show that the posterior mean is a weighted average of the prior mean and the sample mean, in which the prior mean gets β votes and the sample mean gets n votes; this identifies the prior sample size in this model as $n_0 = \beta$. *[10 points]*

(j) Suppose that (as was true in the RAND investigation), before this study was conducted, not much was known external to the data set about θ; this suggests a diffuse prior in which the prior sample size β is small, for example $\Gamma(0.01, 0.01)$. Use `R` to plot the prior, likelihood and posterior on the same graph with this prior and the data set given in this problem. (*Hint:* Be careful with the parameterization of the Gamma distribution in `R`.) Compute the posterior mean and SD (to get the posterior SD, use the variance expression in (**??**)), and compare with the MLE and its standard error in part (f). Use the `qgamma` function in `R` to compute a 95% posterior interval for θ, and compare with Your approximate interval based on the MLE; briefly discuss any differences You find between the maximum-likelihood and Bayesian answers. *[40 points]*