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

AMS 131: Take-Home Test 2

Target due date: Wed 23 Aug 2017 *[520 total points]*

1. *[70 total points]* (the Exchange Paradox) You're playing the following game against an opponent, with a referee also taking part. The referee has two envelopes (numbered 1 and 2 for the sake of this problem, but when the game is played the envelopes have no markings on them), and (without you or your opponent seeing what she does) she puts $\$m$ in envelope 1 and $\$2\,m$ in envelope 2 for some $m > 0$ (treat m as continuous in this problem even though in practice it would have to be rounded to the nearest dollar or penny). You and your opponent each get one of the envelopes at random. You open your envelope secretly and find $\$x$ (your opponent also looks secretly in his envelope), and the referee then asks you if you want to trade envelopes with your opponent. You reason that if you trade, you will get either $\$\frac{x}{2}$ or $\$2\,x$, each with probability $\frac{1}{2}$. This makes the expected value of the amount of money you'll get if you trade equal to $\left(\frac{1}{2}\right)\left(\frac{\$x}{2}\right) + \left(\frac{1}{2}\right)(\$2\,x) = \frac{\$5x}{4}$, which is greater than the $\$x$ you currently have, so you offer to trade. The paradox is that your opponent is capable of making exactly the same calculation. How can the trade be advantageous for both of you?

The point of this problem is to demonstrate that the above reasoning is flawed from a Bayesian point of view; the conclusion that trading envelopes is always optimal is based on the assumption that there's no information obtained by observing the contents of the envelope you get, and this assumption can be seen to be false when you reason in a Bayesian way. At a moment in time before the game begins, let $p(m)$ be your prior distribution on the amount of money M the referee will put in envelope 1, and let X be the amount of money you'll find in your envelope when you open it (when the game is actually played, the observed x, of course, will be data that can be used to decrease your uncertainty about M).

(a) Explain why the setup of this problem implies that $P(X = m|M = m) = P(X = 2\,m|M = m) = \frac{1}{2}$, and use this to show that

$$P(M = x|X = x) = \frac{p(x)}{p(x) + p\left(\frac{x}{2}\right)} \quad \text{and} \quad P\Big(M = \frac{x}{2}\Big|\, X = x\Big) = \frac{p\left(\frac{x}{2}\right)}{p(x) + p\left(\frac{x}{2}\right)}\,. \tag{1}$$

Demonstrate from this that the expected value of the amount Y of money in your opponent's envelope, given than you've found $\$x$ in the envelope you've opened, is

$$E(Y|X = x) = \frac{p(x)}{p(x) + p\left(\frac{x}{2}\right)}(2\,x) + \frac{p\left(\frac{x}{2}\right)}{p(x) + p\left(\frac{x}{2}\right)}\left(\frac{x}{2}\right)\,. \tag{2}$$

[20 points]

(b) Suppose that for you in this game, money and utility coincide (or at least suppose that utility is linear in money for you with a positive slope). Use Bayesian decision theory, through the principle of maximizing expected utility, to show that you should offer to trade envelopes only if

$$p\Big(\frac{x}{2}\Big) < 2\,p(x)\,. \tag{3}$$

If you and two friends (one of whom would serve as the referee) were to actually play this game with real money in the envelopes, it would probably be the case that small amounts of money are more likely to be chosen by the referee than big amounts, which makes it interesting to explore condition (3) for prior distributions that are decreasing (that is, for which $p(m_2) < p(m_1)$ for $m_2 > m_1$). Make a sketch of what condition (3) implies for a decreasing p. One possible example of a continuous decreasing family of priors on M is the *exponential* distribution indexed by the parameter λ, which represents the reciprocal of the mean of the distribution. Identify the set of conditions in this family of priors, as a function of x and λ, under which it's optimal for you to trade. Does the inequality you obtain in this way make good intuitive sense (in terms of both x and λ)? Explain briefly. *[40 points]*

(c) Looking carefully at the correct argument in paragraph 2 of this problem, identify precisely the point at which the argument in the first paragraph breaks down, and specify what someone who believes the argument in paragraph 1 is implicitly assuming about the prior distribution $p(m)$. *[10 points]*

2. *[210 total points]* (practice with joint, marginal and conditional densities) This is a toy problem designed to give you practice in working with a number of the concepts we've examined; in a course like this, every now and then you have to stop looking at real-world problems and just work on technique (it's similar to classical musicians needing to practice scales in addition to actual pieces of symphonic or chamber music).

Suppose that the continuous random vector $\boldsymbol{X} = (X_1, X_2)$ has PDF given by

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \left\{ \begin{array}{cc} 4\,x_1\,x_2 & \text{for } 0 < x_1 < 1, 0 < x_2 < 1 \\ 0 & \text{otherwise} \end{array} \right\} \tag{4}$$

in which $\boldsymbol{x} = (x_1, x_2)$, and define the random vector $\boldsymbol{Y} = (Y_1, Y_2)$ with the transformation ($Y_1 = X_1, Y_2 = X_1\,X_2$).

(a) Are X_1 and X_2 independent? Present any relevant calculations to support your answer. *[10 points]*

(b) Either work out the correlation $\rho(X_1, X_2)$ between X_1 and X_2 or explain why no calculation is necessary in correctly identifying the value of ρ. *[10 points]*

(c) Sketch the set S of possible $\boldsymbol{X}$ values and the image T of S under the transformation from $\boldsymbol{X}$ to $\boldsymbol{Y}$, and show that the joint distribution of $\boldsymbol{Y} = (Y_1, Y_2)$ is

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \left\{ \begin{array}{cc} 4\,\frac{y_2}{y_1} & \text{for } 0 < y_1 < 1, 0 < y_2 < y_1 < 1 \\ 0 & \text{otherwise} \end{array} \right\}, \tag{5}$$

in which $\boldsymbol{y} = (y_1, y_2)$. Verify your calculation by demonstrating that $\iint_T f_{\boldsymbol{Y}}(\boldsymbol{y})\,d\boldsymbol{y} = 1$. *[50 points]*

(d) Work out

(i) the marginal distributions for Y_1 and Y_2, sketching both distributions and checking that they both integrate to 1;

(ii) the conditional distributions $f_{Y_1\,|\,Y_2}(y_1\,|\,y_2)$ and $f_{Y_2\,|\,Y_1}(y_2\,|\,y_1)$, checking that they each integrate to 1; and

(iii) the conditional expectations $E(Y_1 \,|\, Y_2)$ and $E(Y_2 \,|\, Y_1)$; and

(iv) the conditional variances $V(Y_1 \,|\, Y_2)$ and $V(Y_2 \,|\, Y_1)$. (*Hint:* recall that the variance of a random variable W is just $E\,(W^2) - [E(W)]^2$.)

[120 points]

(e) Are Y_1 and Y_2 independent? Present any relevant calculations to support your answer. *[10 points]*

(f) Either work out the correlation $\rho(Y_1, Y_2)$ between Y_1 and Y_2 or explain why no calculation is necessary in correctly identifying the value of ρ. *[10 points]*

3. *[100 total points]* (moment-generating functions) Distributions may in general be skewed, but there may be conditions on their parameters that make the skewness get smaller or even disappear. This problem uses moment-generating functions (MGFs) to explore that idea for two important discrete distributions, the Binomial and the Poisson.

(a) We saw in class that if $X \sim \text{Binomial}(n, p)$, for $0 < p < 1$ and integer $n \geq 1$, then the MGF of X is given by

$$\psi_X(t) = \left[p\, e^t + (1-p)\right]^n \,. \tag{6}$$

for all real t, and we used this to work out the first three moments of X (note that the expression for $E\,(X^3)$ is only correct for $n \geq 3$):

$$E(X) = n\,p\,, \quad E\left(X^2\right) = n\,p[(1 + (n-1)p]\,, \tag{7}$$

$$E\left(X^3\right) = n\,p[1 + (n-2)(n-1)p^2 + 3\,(n-1)p]\,, \tag{8}$$

from which we also found that $V(X) = n\,p(1-p)$. Show that the above facts imply that

$$\text{skewness}(X) = \frac{1 - 2\,p}{\sqrt{n\,p(1-p)}}\,. \tag{9}$$

Under what condition on p, if any, does the skewness vanish? Under what condition on n, if any, does the skewness tend to 0? Explain briefly. *[30 points]*

(b) In our brief discussion of stochastic processes we encountered the *Poisson* distribution: if $Y \sim \text{Poisson}(\lambda)$, for $\lambda > 0$, then the PF of Y is

$$f_Y(y) = \left\{ \begin{array}{cc} \frac{\lambda^y\, e^{-\lambda}}{y!} & \text{for } y = 0, 1, \ldots \\ 0 & \text{otherwise} \end{array} \right\}\,. \tag{10}$$

(i) Use this to show that for all real t the MGF of Y is

$$\psi_Y(t) = e^{\lambda\left(e^t - 1\right)}\,. \tag{11}$$

[10 points]

(ii) Use $\psi_Y(t)$ to compute the first three moments of Y, the variance of Y and the skewness of Y. Under what condition on λ, if any, does the skewness either disappear or tend to 0? Explain briefly. *[60 points]*

4. *[140 total points]* (archaeology) 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 $\{y_i, i = 1, \ldots, n\}$ denote a sample of such distances above P at a random set of locations, the model $\boxed{(Y_i|\theta) \stackrel{\text{IID}}{\sim} \text{Uniform}(0, \theta) \; (*)}$ 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.

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

$$p_{Y_i}(y_i \,|\, \theta) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{if } 0 \le y_i \le \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta} \, I\left(0 \le y_i \le \theta\right) , \tag{12}$$

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

(a) Briefly explain why the statement $\{0 \le y_i \le \theta \text{ for all } i = 1, \ldots, n\}$ is equivalent to the statement $\{m = \max(y_1, \ldots y_n) \le \theta\}$, and use this to show that the joint distribution of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ in this model is

$$f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = \frac{I(m \le \theta)}{\theta^n} \, . \tag{13}$$

[20 points]

(b) Letting the observed values of $(Y_1, \ldots, Y_n)$ be $\boldsymbol{y} = (y_1, \ldots, y_n)$, an important object in both frequentist and Bayesian inferential statistics is the *likelihood function* $\ell(\theta \,|\, \boldsymbol{y})$, which is obtained from the joint distribution of $(Y_1, \ldots, Y_n)$ simply by

(1) thinking of $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$ as a function of θ for fixed $\boldsymbol{y}$, and

(2) multiplying by an arbitrary positive constant c:

$$\ell(\theta \,|\, \boldsymbol{y}) = c \, f_{\boldsymbol{Y}}(\boldsymbol{y}) \, . \tag{14}$$

Using this terminology, in part (a) you showed that the likelihood function in this problem is $\ell(\theta \,|\, \boldsymbol{y}) = \theta^{-n} I(\theta \ge m)$, where m is the largest of the y_i values. Both frequentists and Bayesians are interested in something called the *maximum likelihood estimator* (MLE) $\hat{\theta}_{\text{MLE}}$, which is the value of θ that makes $\ell(\theta \,|\, \boldsymbol{y})$ as large as possible.

(i) Make a rough sketch of the likelihood function, and use your sketch to show that the MLE in this problem is $\hat{\theta}_{\text{MLE}} = m = \max(y_1, \ldots y_n)$. *[20 points]*

(ii) Maximization of a function is usually accomplished by setting its first derivative to 0 and solving the resulting equation. Briefly explain why that method won't work in finding the MLE in this case. *[10 points]*

(c) A positive quantity W follows the *Pareto* distribution (written $W \sim \text{Pareto}(\alpha, \beta)$) if, for parameters $\alpha, \beta > 0$, it has density

$$f_W(w) = \left\{ \begin{array}{cc} \alpha \, \beta^\alpha \, w^{-(\alpha+1)} & \text{if } w \ge \beta \\ 0 & \text{otherwise} \end{array} \right\} . \tag{15}$$

This distribution has mean $\frac{\alpha\beta}{\alpha-1}$ (if $\alpha > 1$) and variance $\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ (if $\alpha > 2$).

(i) For frequentists the likelihood function is just a function $\ell(\theta \,|\, \boldsymbol{y})$, but for Bayesians it can be regarded as an un-normalized density function for θ. Show that, from this point of view, the likelihood function in this problem corresponds to the Pareto$(n-1, m)$ distribution. *[10 points]*

(ii) Bayes's Theorem for a one-dimensional continuous unknown (such as θ in this situation) says that the conditional density $f_{\Theta \,|\, \boldsymbol{Y}}(\theta \,|\, \boldsymbol{y})$ for θ given $\boldsymbol{Y} = \boldsymbol{y}$ — which is called the *posterior distribution* for θ given the data — is a positive (normalizing) constant c times a PDF $f_\Theta(\theta)$ — called the *prior distribution* for θ — that captures any available information about θ external to the data set, times the likelihood distribution $\ell(\theta \,|\, \boldsymbol{y})$:

$$\begin{array}{ccccccc} f_{\Theta \,|\, \boldsymbol{Y}}(\theta \,|\, \boldsymbol{y}) & = & c & \cdot & f_\Theta(\theta) & \cdot & \ell(\theta \,|\, \boldsymbol{y}) \\ \left(\begin{array}{c}\text{posterior}\\ \text{distribution}\end{array}\right) & = & \left(\begin{array}{c}\text{normalizing}\\ \text{constant}\end{array}\right) & \cdot & \left(\begin{array}{c}\text{prior}\\ \text{distribution}\end{array}\right) & \cdot & \left(\begin{array}{c}\text{likelihood}\\ \text{distribution}\end{array}\right) \end{array} \tag{16}$$

The posterior distribution is the goal of a Bayesian inferential analysis: it summarizes *all* available information, both external to and internal to your data set. Show that if the prior distribution for θ in this problem is taken to be (15), under the model $(*)$ above the posterior distribution is $f_{\Theta \,|\, \boldsymbol{Y}}(\theta \,|\, \boldsymbol{y}) = \text{Pareto}\,[\alpha + n, \max(\beta, m)]$. (Bayesian terminology: Note that what just happened was that the product of two Pareto distributions (prior, likelihood) is another Pareto distribution (posterior); a prior distribution that makes this happen is called *conjugate* to the likelihood in the model.) *[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: $y = (y_1, \ldots, y_n) = (2.8, 1.7, 1.0, 5.1, 3.7, 1.5, 4.3, 2.0, 3.2, 2.1, 0.4)$. Prior information equivalent to a Pareto distribution specified by the choice $(\alpha, \beta) = (2.5, 4)$ was available.

(i) Plot the prior, likelihood, and posterior distributions arising from this data set on the same graph, explicitly identifying the three curves. *[30 points]*

(ii) Work out the posterior mean and SD (square root of the posterior variance), and use them to complete the following sentence:

> *On the basis of this prior and data information, the θ value for this species of fossil ammonite is about _____, give or take about _____.*

[30 points]