18 Jan 2013

eBay/Google short course: Problem set 2

1. (the Exchange Paradox) You are 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 \$$2m$ 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 her 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 \$$2x$, each with probability $\frac{1}{2}$. This makes the expected value of the amount of money you'll get if you trade equal to $\left(\frac{1}{2}\right)\left(\frac{\$x}{2}\right) + \left(\frac{1}{2}\right)(\$2x) = \frac{\$5x}{4}$, 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 is 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 will 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 $P(X = m|M = m) = P(X = 2m|M = m) = \frac{1}{2}$, and use this to show that

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

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

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

(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

$$p\left(\frac{x}{2}\right) < 2p(x). \tag{3}$$

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, with density (4) below, indexed by the parameter λ which represents 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.

Extra credit: 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 p.

2. (exchangeability) (a) Suppose Y_1 and Y_2 are identically distributed Bernoulli random variables with success probability $0 < \theta < 1$. Show that independence of Y_1 and Y_2 implies exchangeability but not conversely. The simplest way to do this is to specify their joint distribution by making a 2×2 table cross-tabulating Y_1 against Y_2, labeling all of the probabilities symbolically. What does this table have to look like in terms of θ if Y_1 and Y_2 are independent? What about when they're exchangeable? (In the latter case you'll have to choose a new symbol for some of the relevant probabilities.)

Extra credit: See if you can quantify how far away from independence Y_1 and Y_2 can be (in some sense of distance in the space of possible joint distributions) and still be exchangeable.

(b) Can you give another simple example, involving a comparison of random sampling with and without replacement from a finite population, of a set of random variables that are exchangeable but not independent? Explain briefly.

3. (Bayesian conjugate inference in the exponential distribution) In a consulting project that one of my Ph.D. students and I worked on at the University of Bath in England before I came to Santa Cruz, a researcher from the Department of Electronic and Electrical Engineering (EEE) at Bath wanted help in analyzing some data on failure times for a particular kind of metal wire (in this problem, failure time was defined to be the number of times the wire could be mechanically stressed by a machine at a given point along the wire before it broke). The $n = 14$ raw data values y_i in one part of his experiment, arranged in ascending order, were

495 541 1461 1555 1603 2201 2750 3468 3516 4319 6622 7728 13159 21194

Probably the simplest model for failure time data is the *exponential* distribution $\mathcal{E}(\lambda)$:

$$(y_i|\lambda) \stackrel{\text{IID}}{\sim} p(y_i|\lambda) = \left\{ \begin{array}{cc} \frac{1}{\lambda}\exp(-\frac{y_i}{\lambda}) & y_i > 0 \\ 0 & \text{otherwise} \end{array} \right\} \qquad (4)$$

for some $\lambda > 0$. (**NB** This distribution can be parameterized either in terms of λ or $\frac{1}{\lambda}$; whenever it occurs in print you need to be careful which parameterization is in use.)

(a) To see if this model fits the data above, you can make an *exponential probability plot*, analogous to a Gaussian quantile-quantile plot to check for normality. In fact the idea works for more or less any distribution: you plot

$$y_{(i)} \quad \text{versus} \quad F^{-1}\left(\frac{i-0.5}{n}\right), \qquad (5)$$

where $y_{(i)}$ are the sorted y values and F is the CDF of the distribution (the 0.5 is in there to avoid problems at the edges of the data). In so doing you're graphing the data values against an approximation of *what you would have expected for the data values if the CDF of the y_i really had been F*, so the plot should resemble the 45° line if the fit is good.

(i) Show that the inverse CDF of the $\mathcal{E}(\lambda)$ distribution (parameterized as in equation (4)) is given by

$$F_Y(y|\lambda) = p \iff y = F^{-1}(p) = -\lambda\log(1-p). \qquad (6)$$

(ii) To use equation (6) to make the plot we need a decent estimate of λ. Show that the maximum likelihood estimate of λ in this model is $\hat{\lambda}_{\text{MLE}} = \bar{y}$, the sample mean, and use this (in `Maple`, or freehand, or with whatever other software you might like) to make an exponential probability plot of the 14 data values above. Informally, does the exponential model appear to provide a good fit to the data? Explain briefly.

(b) (i) Show that the exponential sampling model (4) is a member of the one-parameter exponential family, and use this to show that the conjugate family for the $\mathcal{E}(\lambda)$ likelihood (parameterized as in (4)) is the set of *Inverse Gamma* distributions $\Gamma^{-1}(\alpha,\beta)$ for $\alpha > 0, \beta > 0$ (**NB** $W \sim \Gamma^{-1}(\alpha,\beta)$ just means that $\frac{1}{W} \sim \Gamma(\alpha,\beta)$; see Table A.1 from Appendix A in Gelman et al., a copy of which will be distributed in class, for details):

$$\lambda \sim \Gamma^{-1}(\alpha,\beta) \quad \iff \quad p(\lambda) = \left\{ \begin{array}{cc} \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{-(\alpha+1)}\exp\left(-\frac{\beta}{\lambda}\right) & \lambda > 0 \\ 0 & \text{otherwise} \end{array} \right\}. \qquad (7)$$

(ii) By directly using Bayes' Theorem (and ignoring constants), show that the prior-to-posterior updating rule in this model is

$$\left\{ \begin{array}{ccc} \lambda & \sim & \Gamma^{-1}(\alpha,\beta) \\ (Y_i|\lambda) & \stackrel{\text{IID}}{\sim} & \mathcal{E}(\lambda) \end{array} \right\} \Longrightarrow (\lambda|y) \sim \Gamma^{-1}(\alpha+n, \beta+n\bar{y}). \qquad (8)$$

(iii) It turns out that the mean and variance of the $\Gamma^{-1}(\alpha,\beta)$ distribution are $\frac{\beta}{\alpha-1}$ and $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$, respectively (as long as $\alpha > 2$). Use this to write down an explicit formula which shows that the posterior mean is a weighted average of the prior and sample means, and deduce from this formula that $n_0 = (\alpha - 1)$ is the prior effective sample size. Note also from the formula for the likelihood in this problem that, when thought of as a distribution in λ, it's equivalent to a constant times the $\Gamma^{-1}(n-1, n\bar{y})$ distribution.

(c) The guy from EEE has prior information from another experiment he judges to be comparable to this one: from this other experiment the prior for λ should have a mean of about $\mu_0 = 4500$ and an SD of about $\sigma = 1800$.

(i) Show that this corresponds to a $\Gamma^{-1}(\alpha_0, \beta_0)$ prior with $(\alpha_0, \beta_0) = (8.25, 32625)$, and therefore to a prior sample size of about 7.

(ii) The next step is to work on filling in the entries in the following table:

		Likelihood		
	Prior	Maximizing	Integrating	Posterior
Mean/Estimate	4500			
SD/SE	1800			

Show that the Fisher information provided by the MLE in this model is

$$\hat{I}\left(\hat{\lambda}_{\text{MLE}}\right) = \frac{n}{\bar{y}^2}, \tag{9}$$

so that a large-sample standard error for the MLE is

$$\widehat{SE}\left(\hat{\lambda}_{\text{MLE}}\right) = \frac{\bar{y}}{\sqrt{n}}. \tag{10}$$

In the "Likelihood Maximizing" column put the numerical values of the MLE and its standard error; in the "Likelihood Integrating" column put the mean and SD of the likelihood function, interpreted as the $\Gamma^{-1}(n-1, n\bar{y})$ distribution; and in the "Posterior" column put the posterior mean and SD using the (α_0, β_0) and data values above. By examining the formulas for the relevant quantities, show that the discrepancies between the "Likelihood Maximizing" and "Likelihood Integrating" columns in this table will diminish as n increases.

(iii) What kind of compromise, if any, gives rise to the posterior SD as a function of the prior and likelihood SDs, at least approximately? Explain briefly.

(iv) Make a plot with `Maple`, or an approximate freehand sketch, of the prior, likelihood ($\Gamma^{-1}(n-1, n\bar{y})$), and posterior distributions on the same graph, and summarize what all of this has to say about the failure times of the metal wire samples with which the problem began.

(d) (comparing Bayesian and [large-sample] maximum likelihood interval estimates) From the Fisher information calculation above, an approximate 95% interval estimate for λ in this model based on the (large-sample) likelihood approach has the form

$$\bar{y} \pm 1.96 \frac{\bar{y}}{\sqrt{n}}. \tag{11}$$

By using the numerical integration features in `Maple` I've computed the endpoints of 95% central intervals based both on the posterior distribution in (c) and the likelihood distribution $\Gamma^{-1}(n-1, n\bar{y})$, obtaining $(3186, 7382)$ and $(3369, 10201)$, respectively. (**NB** (a, b) is a $100(1-\alpha)\%$ central interval for a real-valued parameter θ with respect to an inferential density $p(\theta)$ if $\int_{-\infty}^{a} p(\theta)\, d\theta = \int_{b}^{\infty} p(\theta)\, d\theta = \frac{\alpha}{2}$.) Compute the (large-sample) likelihood interval (11) above on this dataset and explain briefly why it's not directly comparable to the 95% posterior interval. In what way does your plot of the likelihood function in (c) suggests that the central likelihood interval might be better than interval (11) for a value of n as small as the one in this problem? Explain briefly.

Extra credit: Compute the predictive distribution for the next observation Y_{n+1} given $y = (y_1, \ldots, y_n)$ in model (8). Apply this to the data set on page 2 with the largest observation (21194) set aside, using a diffuse Inverse Gamma prior (e.g., pick that member of the Inverse Gamma family that has mean 1 and precision ϵ for some small ϵ like 0.001, by analogy with the $\Gamma(\epsilon, \epsilon)$ prior), and compute a numerical measure of how surprising this observation is under the exponential model. How strongly, if at all, do your calculations call into question this model for these data? Explain briefly.