Course 52558: Problem Set 1 Solution

1. **A Thought Experiment**. Let θ be the true proportion of men in Israel over the age of 40 with hyper-tension.

   (a) Though you may have little or no expertise in this area, use your social knowledge and common sense to give an initial point estimate (single value) of θ.

   Solution:
   As an example lets guess that 20% of the population has hypertension so that our point "estimate" for θ is 0.2.

   (b) Now suppose that in a properly designed survey, of the first 5 randomly selection men, 4 are hypertensive. How does this information effect your initial estimate of θ?.

   Solution:
   Naturally, having observed that 80% of the sample is hyper-tensive causes us to suspect that our initial estimate may have been low. Yet, due to the small sample we still believe that it is not the case the most of the population is hyper-tensive and might update our belief to, say, one third.

   (c) Finally suppose that at the survey's completion, 400 of 1000 men have emerged as hypertensive. Now what is your estimate of θ?

   Solution:
   A properly selected random sample of 1000 is significant enough to convince us that the true value of θ is close to 0.4. Although this is higher than our original estimate, it is still less than the majority and we are willing to let the sample surprise us to that extent.

   (d) What guidelines for statistical inference do your answers suggest?

   Solution:
   The above suggests the following sensible data analysis steps

   i. Plausible starting estimates
   ii. Gradual revision of beliefs
   iii. Eventual convergence to the data

2. **Relationship between posterior and prior mean and variance**.

   (a) Show that for any continuous random variables X and Y

$$E(X) = E(E(X \mid Y))$$

(Note that a similar proof can be used for the discrete case)

Solution:

$$\begin{aligned} E(X) &= \int\int Xp(X,Y)dXdY = \int\int Xp(X \mid Y)dup(Y)dv \\ &= \int E(X \mid Y)p(y)dy = E(E(X \mid Y)) \end{aligned}$$

(b) Show that for any random variables X and Y

$$\text{var}(X) = E(var(X \mid Y)) + var(E(X \mid Y))$$

Solution:

For variety, I will prove this one for the discrete case

$$\begin{aligned} &E(var(X \mid Y) + var(E(X \mid Y)) \\ &\quad= E(E(X^2 \mid Y) - (E(X \mid Y))^2) + E((E(X \mid Y))^2) - (E(E(X \mid Y)))^2 \\ &\quad= E(X^2) - E((E(X \mid Y))^2) + E((E(X \mid Y))^2) - (E(X))^2 \\ &\quad= E(X^2) - (E(X))^2 = var(X) \end{aligned}$$

(c) Let y denote the observed data. We assume y was generated from $p(y \mid \theta)$, where θ, the parameters governing the sampling of y are random and distributed according to $p(\theta)$. Use the above and describe (i.e. understand the equation and then put into words) the relationship between the mean and variance of the prior $p(\theta)$ and and the posterior $p(\theta \mid y)$.

Solution:

Plugging θ and y into the first equation gives us

$$E(\theta) = E(E(\theta \mid y))$$

which means that the prior mean over the parameters is the overage over all possible posterior means over the distribution of possible data. This is in the opposite direction of the way we are used to think of priors and posterior but in fact matches our intuition of what the prior should capture. The second equation

$$var(\theta) = E(var(\theta \mid y)) + var(E(\theta \mid y))$$

is interesting because it means that the posterior variance (left term of right hand-side) is *on average* smaller than the prior variance (left-side) by an amount that depends on the variation of the posterior mean. Importantly, the greater the variation of the posterior mean, the greater our potential for reducing our uncertainty about θ.

3. **Posterior of a Poisson Distribution**. Suppose that X is the number of pregnant woman arriving at a particular hospital to deliver babies in a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest adopting a Poisson likelihood

$$p(x \mid \theta) = \frac{e^{-\theta}\theta^x}{x!}, x \in \{0, 1, 2, \ldots\}, \theta > 0$$

To provide support on the positive real line and reasonable flexibility we suggest a Gamma $G(\alpha, \beta)$ distribution prior

$$p(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \theta > 0, \alpha > 0, \beta > 0$$

where $\Gamma()$ is a continuous generalization of the factorial function so that $\Gamma(c) = c\Gamma(c-1)$. α, β are the parameters of this prior, or the hyperparameters of the model. The Gamma distribution has mean $\alpha\beta$ and variance $\alpha\beta^2$.

Show that the posterior distribution $p(\theta \mid x)$ is also Gamma distributed. Determine its parameters α and β.

Solution:
Although the question was phrased for a univariate x, for generality, in the solution x will be a vector of observations x_i, each of which follows the Poisson distribution. In this case, we have

$$p(x \mid \theta) \propto \theta^{\sum_i x_i} e^{-n\theta}$$

Together with the prior we can the write

$$\begin{aligned} p(\theta \mid x) &\propto \theta^{\sum_i x_i} e^{-n\theta} \theta^{\alpha-1} e^{-\theta/\beta} \\ &= \theta^{\sum_i x_i + \alpha - 1} e^{-\theta(n+\frac{1}{\beta})} \end{aligned}$$

This is an unnormalized Gamma distribution so that

$$\theta \mid x \sim Gamma(\sum_i x_i + \alpha, \frac{1}{n+\frac{1}{\beta}} = \frac{\beta}{n\beta+1})$$

4. **Posterior of the Poisson Model**.

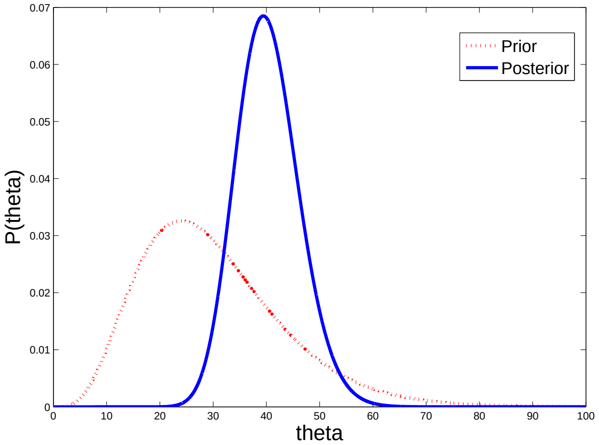
   In this question we will use Matlab/R to explore the Poisson model with the Gamma prior considered above.

   (a) The Matlab (GammaPrior.m) R (GammaPrior.txt) files in the Code directory of the course web page can be used to plot the Gamma prior. Use this code to investigate different values for α and β. Describe the qualitative behavior of this prior as a function of these parameters and try to explain why they are called 'shape' and 'scale' parameters, respectively.

Solution:
For a fixed α, the larger the value of β the wider the distribution and thus behaves like a stretching or scale parameter. For a fixed β, α determines the form of the distribution. At $\alpha = 1$ we see a singular point where for $alpha <= 1$ the distribution starts with a high value and decays exponentially, and for $alpha > 1$ the distribution is a unimodal one that resembles a normal distribution more as α grows.

(b) Continuing the previous question involving births, assume that in December 2008 we observed $x = 42$ moms arriving at the hospital to deliver babies, and suppose we adopt a Gamma(5,6), which has mean 30 and variance 180, reflecting the hospital's total for the two preceding years. Use Matlab/R to plot the posterior distribution of θ next to its prior. What are your conclusions?

Solution:
In this case we have a single observation of $x = 42$ and our posterior is $Gamma(42 + 5, 6/(1 * 6 + 1))$.



It is clear that the mode of the posterior distribution is attracted by the maximum likelihood mode of 42 and the distribution is more concentrated than the prior since our uncertainty has diminished.

(c) Repeat the above for different values of x. What are your conclusions.
Note that different values of x only effect α and not β so that the most significant change is the location of the mode of the posterior distribution. When closer the maximum likelihood mode to that of the prior, the closer the posterior. Since α effects the mode linearly the posterior mode is at approximately (because of the -1 term) a constant fraction of the way between the prior and the posterior. The variance grows linearly with α and so larger values of x also lead to a wider posterior.

5. **Extinction of Species**. Paleobotanists estimate the moment in the remote past when a given species became extinct by taking cylindrical, vertical core samples well below the earths 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 $\{y_1, \ldots, y_n\}$ denote a sample of such distances above P at a random set of locations, the model

$$(y_i|\theta) \sim \text{Unif}(0, \theta)$$

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 Bayesian inference for θ, and it will be seen that some of our usual intuitions do not quite hold in this case.

(a) Show that the likelihood may be written as

$$l(\theta : y) = \theta^{-n} I(\theta \geq max(y_1, \ldots, y_n))$$

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

Solution:

$$\begin{aligned} l(\theta : y) &= \prod_i p(y_i \mid \theta) = \prod_i \frac{1}{\theta} I(0 < y_i \leq \theta) \\ &= \theta^{-n} \prod_i I(y_i \leq \theta) = \theta^{-n} I(\theta \geq max(y_1, \ldots, y_n)) \end{aligned}$$

where we also used the fact that all y_i, by the experiment design are guaranteed to be positive.

(b) The Pareto(α, β) distribution has density

$$p(\theta) = \begin{cases} \alpha\beta^\alpha\theta^{-(\alpha+1)} & \theta \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha, \beta > 0$. The Pareto distribution has mean $\frac{\alpha\beta}{\alpha-1}$ for $\alpha > 1$ and a variance of $\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $alpha > 2$.

With the likelihood viewed as a constant multiple of a density for θ, show that the likelihood corresponds to the Pareto$(n1, m)$ distribution. Now let the prior for θ be taken to be Pareto(α, β) and derive the posterior distribution $p(\theta|y)$. Is the Pareto conjugate to the uniform?

Solution:

We define $m = (y_1, \ldots, y_n)$ (this was supposed to be part of the question but was omitted by mistake). The likelihood can then be written as

$$l(\theta : y) = \theta^{-n} I(m \leq \theta) \propto \theta^{-[(n-1)+1]}(n-1)m^{n-1}1(m \leq \theta)$$

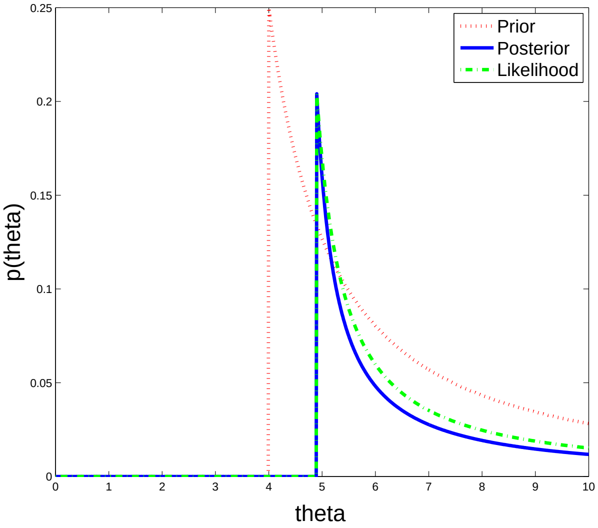
which is, by definition, a Pareto distribution with parameters $n-1$ and m. Together with a Pareto(α, β) prior we have

$$\begin{aligned} p(\theta \mid y) &\propto \theta^{-n} I(m \leq \theta) \alpha \beta^{\alpha} \theta^{-(\alpha+1)} I(\beta \leq \theta) \\ &\propto \theta^{-(n+\alpha+1)} I(max(\beta, m) \leq \theta) \end{aligned}$$

which is a Pareto$(n+\alpha, max(\beta, m))$ distribution and thus the uniform distribution is indeed a conjugate prior.

(c) In an experiment in the Antarctic in the 1980s to study a particular species of fossil ammonite, the following is a linearly rescaled version of the data obtained: $y = (0.4, 1.0, 1.5, 1.7, 2.0, 2.1, 3.1, 3.7, 4.3, 4.9)$. Prior information equivalent to a Pareto prior with $(\alpha, \beta) = (2.5,\ 4)$ was available. Plot the prior, likelihood, and posterior distributions arising from this data set on the same graph, and briefly discuss what this picture implies about the updating of information from prior to posterior in this case.

Solution:
In our case $n = 10$ and $m = max(y_1, \dots, y_10) = 4.9$ and $max(m, \beta) = 4.9$. Thus the likelihood is a Pareto$(9, 4.9)$ and the posterior is a Pareto$(12.5, 4.9)$. Using Matlab, a call to the generalized Pareto distribution $\mathit{gppdf}(\alpha, \beta, \beta)$ can be used to plot the Pareto distribution.



(d) Make a table summarizing the mean and standard deviation for the prior, likelihood and posterior distributions, using the (α, β) choices and the data in part (d) above. In Bayesian updating the posterior mean is often a weighted average of the prior mean and the likelihood mean (with positive weights), and the posterior standard deviation is typically smaller than either the prior or likelihood standard deviations. Are each of these behaviors true in this case? Explain briefly.

Solution:
For the prior,likelihood and posterior distribution, the shape (α) parameter of the Pareto distribution is greater than 2 so that both the mean and standard deviation are finite (see equations above). Specifically we have

	Prior	Likelihood	Posterior
Mean	6.667	5.513	5.326
STD	5.9628	0.6945	0.4649

In this case the posterior standard deviation, as expected is smaller than both that of the prior and the likelihood distribution. However, non-typically, the mean of the posterior is further away from the prior mean than the likelihood mean. This is a result of the unique nature of the Pareto distribution and can be seen directly by plugging, for a fixed β, the appropriate shape parameter into the mean equation of the Pareto distribution.