PHP 2530: BAYESIAN STATISTICAL METHODS
HOMEWORK I SOLUTIONS

NICK LEWIS

Problem 1

Basu's elephant example was described in class. In this example the circus statistician uses probability sampling to obtain an estimate of the total weight. Assume that the sample size is 1, and use the statistician estimator composed of the weight of the sampled elephant, multiplied by the inverse probability of an elephant to be sampled. Calculate the variance of this estimator (This estimator and its variance are also known as the Horvitz–Thompson estimators, and as we saw in class can produce some illogical results.)

Solution:

Define $E_w = \{w_1, w_2, \ldots, w_N\}$ to be the weights of the N elephants owned by the circus. Let w_j be the weight of elephant j, and p_j be the probability of choosing elephant j into our sample. Let A be a subset of distinct elements of E_w, where the probability that an element $j \in E_w$ is chosen to be in A is denoted p_j. The elephants are sampled without replacement so each element of A is unique. Also, p_j is a known quantity, and the weight w_j is considered to be fixed. The randomness involved in the Horvitz Thompson estimator arises from choosing the sample, not the sample itself (Recall in the example given in class you're trying to sample the weight of a elephant known to have the median weight).

Our Horvitz-Thompson Estimator takes the form $\hat{T} = \sum_{j \in A} \frac{w_j}{p_j}$, in other words the summation of the weights divided by their selection probabilities in the chosen sample. Our sample size for this problem is 1 (the language in the problem can be slightly confusing. It's not that we have one sample, but that out of the N elephants we only choose 1 to weigh) so this estimator reduces to the form $\hat{T} = \frac{w_j}{p_j}$, with $A = \{j\}$, and $1 \leq j \leq N$. Note that j here simply refers to the elephant chosen from E_w. You'll notice that $\hat{T}$ by itself is not random so taking the variance of this term will result in 0 since at this point it is a constant.

We need to think slightly harder about what to do. One thing is to recall the frequentist principle of repeated sampling. If we repeated our sampling procedure many times we would expect different samples to be chosen each time since each elephant has a nonzero probability of being selected. To add this randomness into the estimator we introduce a random variable δ_j^A which takes the form:

$$\delta_j^A = \begin{cases} 1 & \text{if } j \in A \\ 0 & \text{if } j \notin A \end{cases}$$

We can see that $\delta_j^A \sim Bernoulli(p_j)$. Using this we can calculate the variance of $\hat{T}$.

$$Var(\hat{T}) = Var(\sum_{j=1}^{N} \delta_j^A \frac{w_j}{p_j})$$
$$= \sum_{j=1}^{N} \frac{w_j^2}{p_j^2} Var(\delta_j^A) + 2 \sum_{j \neq i} \frac{w_j w_i}{p_j p_i} Cov(\delta_j^A, \delta_i^A)$$

$Cov(\delta_i^A, \delta_j^A) = E[\delta_j^A \delta_i^A] - E[\delta_i^A]E[\delta_j^A] = P(\delta_j^A = 1, \delta_i^A = 1) - p_i p_j = p_{ij} - p_i p_j$. p_{ij} is the probability both i and j are selected into A.

Now note that in the general case where we choose n out of N samples, $\sum_{j=1}^{N} \delta_j^A = n$ gives us that $\sum_{j=1}^{N} p_j = n$ when taking the expectation. This also gives us the identity that

$$\sum_{j \neq i}^{N} p_{ij} = \sum_{j \neq i}^{N} p_{j|i} p_i$$
$$= p_i \sum_{j \neq i}^{N} p_{j|i}$$
$$= p_i \sum_{j \neq i}^{N} E(\delta_j^A \mid \delta_i^A = 1)$$
$$= p_i E(\sum_{j \neq i}^{N} \delta_j^A \mid \delta_i^A = 1)$$
$$= (n-1) p_i$$

$p_{j|i} = Pr(\delta_j^A = 1 \mid \delta_i^A = 1)$. Under this probability, there are only n-1 spaces left within the sample. So we can see for a sample size of 1, the joint probabilities disappear. We can also logically determine this since A only contains one element so the probability of two distinct elements being in A is 0. This simplifies the variance down to:

$$Var(\hat{T}) = \sum_{j=1}^{N} \frac{w_j^2}{p_j}(1 - p_j) - 2 \sum_{j \neq i} w_i w_j$$

Note: One can also reason that the covariance here is nonzero due to the dependence between indicators. If one of them equals 1, then the rest shut off.

Problem 2: BDA 3rd Ed. 1.6

Conditional probability: approximately 1/125 of all births are fraternal twins and 1/300 of births are identical twins. Elvis Presley had a twin brother (who died at birth). What is the probability that Elvis was an identical twin? (You may approximate the probability of a boy or girl birth as $\frac{1}{2}$.)

Solution

We first define our quantity of interest which is:

$$Pr(\text{Identical Twin}|\text{Twin Brother}) = \frac{Pr(\text{Identical Twin \& Twin Brother})}{Pr(\text{Twin Brother})}$$

Since identical implies the twins are either both boys or both girls, the probability in the numerator reduces to $Pr(\text{Both Boys\&Identical Twin}) = \frac{1}{300}\frac{1}{2}$. We can calculate the bottom probability using the total law of probability:

$$\text{Pr(Twin Brother)} = \text{Pr(Identical Twin) Pr(Both Boys | Identical Twin)} + Pr(\text{Fraternal Twin})Pr(\text{Both Boys | Fraternal Twin})$$

Since fraternal implies that both genders need not be the same, we get that
$Pr(\text{Fraternal Twin})Pr(\text{Both Boys} \mid \text{Fraternal Twin}) = \frac{1}{125}\frac{1}{4}$ so we find that

$$\begin{aligned} Pr(\text{Identical Twin}|\text{Twin Brother}) &= \frac{\frac{1}{300}\frac{1}{2}}{\frac{1}{300}\frac{1}{2} + \frac{1}{125}\frac{1}{4}} \\ &= \frac{5}{11} \end{aligned}$$

Problem 3: BDA 3rd Ed. 1.9

Simulation of a queuing problem: a clinic has three doctors. Patients come into the clinic at random, starting at 9 a.m., according to a Poisson process with time parameter 10 minutes: that is, the time after opening at which the first patient appears follows an exponential distribution with expectation 10 minutes and then, after each patient arrives, the waiting time until the next patient is independently exponentially distributed, also with expectation 10 minutes. When a patient arrives, he or she waits until a doctor is available. The amount of time spent by each doctor with each patient is a random variable, uniformly distributed between 15 and 20 minutes. The office stops admitting new patients at 4 p.m. and closes when the last patient is through with the doctor.

(a) Simulate this process once. How many patients came to the office? How many had to wait for a doctor? What was their average wait? When did the office close?

Solution:

To set this problem up, recall that a Poisson Process is a counting process $\{N(t) : t \geq 0\}$ that represents the total number of occurrences or events that have happened up to and including time t, that satisfies the following three properties:

- N(0) = 0, the number of events at the beginning is 0.
- N(t) has independent increments, i.e. for every integer K such that $t_0 = 0 < t_1 < t_2 < \cdots < t_K$, the family $(N((t_j, t_{j+1}])_{j=0}^{K}$ forms a set of independent random variables.
- $N(t) \sim Poisson(\lambda t)$, i.e. the number of events in any interval of length t is a Poisson random variable with parameter λt

It can be shown that the time between arrivals T, the interarrival time, $T \sim Exp(\lambda)$. Here is a quick "derivation." Let T_1 be the time of the first arrival.

$$
\begin{aligned}
Pr(T_1 > t) &= P(N((0,t]) = 0) \\
&= \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \\
&= e^{-\lambda t}
\end{aligned}
$$

The first line says that the probability that the first arrival occurs after time t is equivalent to there being no arrivals from to 0 to t. This gives us $T_1 \sim Exp(\lambda)$. Using independence of increments, one can show that interarrival times are independent and identically distributed.

For our problem we're given that $\lambda = \frac{1}{10}/min$ (follows from the expectation of T, $E(T|\lambda) = \frac{1}{\lambda} = 10$ minutes) and that the clinic is open from 9:00am to 4:00pm, a total of 7 hours or 420 minutes. Setting $t = 420$, a strategy to solve this problem is displayed in the appendix. The results are shown below:

Total Patients	Number of Waiting Patients	Average Wait Time	Closing Time
39	14	13.29 min	429.92 min

We see that the office admitted 39 total patients, 14 had to wait for a doctor, the mean waiting time was 13.29 minutes (about 13 minutes and 17 seconds), and the closing time was 429.92 minutes after opening (this equals about 4:10 PM).

Note: let me make a clear distinction between arrival and interarrival times. Interarrival times are times between arrivals. Arrival times are the times that an individual arrives relative to the starting point. What's more interesting is that our arrival times $A_j = \sum_{k=1}^{j} T_k$ are the sums of independent and identically distributed Exponentials, meaning that $A_j \sim Gamma(j, \lambda)$.

(b) Simulate the process 100 times and estimate the median and 50% interval for each of the summaries in (a).

Solution:

The code to produce the following output is in the appendix.

Summary Statistic	Median	50 % Interval
Total Patients	42	(37, 46)
Number of Waiting Patients	12	(7, 17)
Average Wait Time	6.39 min	(4.93 min, 8.19 min)
Closing Time	4:12pm	(4:05 pm, 4:16pm)

Problem 4: BDA 3rd Ed. 2.4

Predictive distributions: let y be the number of 6's in 1000 independent rolls of a particular real die, which may be unfair. Let θ be the probability that the die lands on '6.' Suppose your prior distribution for θ is as follows:

$$
\begin{aligned}
Pr(\theta = 1/12) &= 0.25 \\
Pr(\theta = 1/6) &= 0.50 \\
Pr(\theta = 1/4) &= 0.25
\end{aligned}
$$

(a) Using the normal approximation for the conditional distributions, $p(y|\theta)$, sketch your approximate prior predictive distribution for y.

Solutions:

If we consider θ to be the probability that the die lands on '6.' we can model y conditional on θ as $y|\theta \sim Bin(n, \theta)$, where $n = 1,000$. Given the large sample size we can then use the normal approximation $y|\theta \sim N(n\theta, n\theta(1-\theta))$. Our prior predictive distribution then takes the form

$$p(y) = \sum_{\theta} p(y|\theta)p(\theta) = N(\frac{250}{3}, \frac{2750}{36})Pr(\theta = 1/12) + N(\frac{500}{3}, \frac{5000}{36})Pr(\theta = 1/6) + N(250, \frac{375}{2})Pr(\theta = 1/4)$$

We plot the probability density function below for reference. This plot is produced by ggplot, and the code to produce is below. This can also be done similarly in python and the code for doing so is shown below as well.

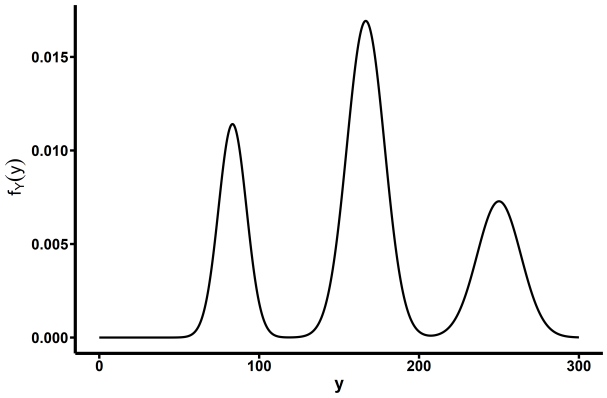

Figure 1. Probability Density Function of Gaussian Mixture Distribution

(b) Give approximate 5%, 25%, 50%, 75%, and 95% points for the distribution of y. (Be careful here: y does not have a normal distribution, but you can still use the normal distribution as part of your analysis.)

Solutions:

Unfortunately there exists no closed form quantile function for the Gaussian Mixture Model. However, there are a multitude of strategies to find the quantiles, some not presented here.

One strategy to find the quantiles is to recognize that the weights (i.e. the prior probability values) reveal how much each individual gaussian contributes to the pdf.

The leftmost Gaussian contributes 25% of the mass, so the 5% quantile is simply the 20% quantile of this Gaussian (i.e. 20% of 25 is 5).

Following similar logic the 25% quantile is directly between the first spike and the second spike (through trial and error I have found that taking the 99.997% of the first Gaussian gives extremely close results).

By symmetry, the 50% quantile is directly in the middle of the second spike.

The 75% quantile is directly between the second and third spikes (taking the 99.96% quantile of the second provides accurate results),

The 95% quantile is given by the 80% quantile of the third spike since the previous two gaussians contribute 75% of the mass, we look for where the third gaussian contributes 20% (i.e. 80% of 25 is 20).

Another equally valid method is to recognize that

$$\int_{-\infty}^{y} p(y')dy' = \int_{-\infty}^{y} \sum_{\theta} p(y'|\theta)p(\theta)dy' = \sum_{\theta} \int_{-\infty}^{y} p(y'|\theta)p(\theta)dy'$$
$$F(y) = \sum_{\theta} F(y|\theta)p(\theta)$$

This gives us a form for the cdf of a mixture model. From here we can use a line-search method to find the values y such that $F(y) - q = 0$, where q is our quantile value.

Regardless of the method chosen to find the quantiles, the resulting quantiles below should be consistent.

5%	25%	50%	75%	95%
76	118	166.67	206.45	261.52

Problem 5: BDA 3rd Ed. 2.7

Noninformative prior densities:

Solution:

(a) For the binomial likelihood, $y \sim Bin(n, \theta)$, show that $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ is the uniform prior distribution for the natural parameter of the exponential family.

Solution:

A single parameter exponential family is a set of probability distributions whose probability density/mass functions can be written in the form

$$f_Y(y|\theta) = e^{\eta(\theta)\cdot T(y) - A(\theta) + B(y)}$$

Recall that the probability mass function for the binomial distribution is $p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$. Using basic algebra this can be rewritten as

$$\begin{aligned} p(y|\theta) &= \binom{n}{y}\theta^y(1-\theta)^{n-y} \\ &= e^{ylog\frac{\theta}{1-\theta} + nlog(1-\theta) + log\binom{n}{y}} \\ &= e^{y\eta(\theta) - nlog(1+e^{\eta(\theta)}) + log\binom{n}{y}} \end{aligned}$$

where $\eta(\theta) = log\frac{\theta}{1-\theta}$. To show that $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ is the uniform prior distribution for the natural parameter of the exponential family, we will do a change of variables.

$$\begin{aligned} p(\eta(\theta)) &\propto \Big|\frac{d\theta}{d\eta}\Big| p(\theta(\eta)) \\ &\propto \Big| -\frac{e^{-\eta}}{(1+e^{-\eta})^2}\Big| \frac{(1+e^{-\eta})^2}{e^{-\eta}} \\ &\propto 1 \end{aligned}$$

which shows the result. Doing the reverse by starting from $p(\eta(\theta)) \propto 1$ is equally valid.

(b) Show that if $y = 0$ or n, the resulting posterior distribution is improper.

Solution:

This part will be easier if we stick to using the posterior of θ. The prior $p(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ is essentially $\theta \sim Beta(0,0)$ (This is technically not a Beta Distribution; just useful shorthand notation for one. This prior is actually called a Haldane Prior), which gives a posterior distribution $\theta|y \sim Beta(y, n-y)$. The integral of the posterior density becomes:

$$\int_0^1 p(\theta|y)d\theta = \begin{cases} \int_0^1 \theta^{-1}(1-\theta)^{n-1}d\theta & \text{if } y = 0 \\ \int_0^1 \theta^{n-1}(1-\theta)^{-1}d\theta & \text{for } y = n \end{cases}$$

It can be shown that the posterior is improper using the binomial theorem, i.e. $(x+y)^n = \sum_{k=0}^n \binom{n}{k} x^k y^{n-k}$. For $y = 0$.

$$\begin{aligned} \theta^{-1}(1-\theta)^{n-1} &= \theta^{-1} \sum_{k=0}^{n-1} \binom{n-1}{k} (-\theta)^k \\ &= \sum_{k=0}^{n-1} \binom{n-1}{k} (-\theta)^{k-1} \\ &= (-\theta)^{n-2} + \binom{n-1}{n-2}(-\theta)^{n-3} + \cdots - \frac{1}{\theta} \end{aligned}$$

We can easily see that when integrating, the polynomial portion will be finite, but $\int_0^1 \frac{1}{\theta} d\theta = \infty$ making the posterior density improper. This argument can be repeated for $y = n$ by using the fact that $\theta = 1 - (1-\theta)$. If one wishes to proceed using the natural parameter formulation, a simple u-substitution followed by partial fraction decomposition will yield the same answer.

Problem 6: BDA 3rd Ed. 2.8

Normal distribution with unknown mean: a random sample of n students is drawn from a large population, and their weights are measured. The average weight of the n sampled students is $\bar{y} = 150$ pounds. Assume the weights in the population are normally distributed with unknown mean θ and known standard deviation 20 pounds. Suppose your prior distribution for θ is normal with mean 180 and standard deviation 40.

(a) Give your posterior distribution for θ. (Your answer will be a function of n.)

Solution:

Let y_j denote the weight of individual j. We're given the following model:

$$\begin{aligned} y|\theta &\sim N(\theta, \sigma^2 = 20^2) \\ \theta &\sim N(\mu = 180, \tau^2 = 40^2) \end{aligned}$$

For the sake of generality we will use the parameters to solve this problem and plug in their values at the end.

$$
\begin{aligned}
p(\theta|y) &\propto \Big(\prod_{j=1}^{n} p(y_j|\theta)\Big)p(\theta) \\
&= \Big(\prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_j-\theta)^2}{2\sigma^2}}\Big) \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \\
&= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \frac{1}{\sqrt{2\pi\tau^2}} e^{-\sum_{j=1}^{n} \frac{(y_j-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\tau^2}}
\end{aligned}
$$

We can simplify this by completing the square inside the exponent.

$$
\begin{aligned}
-\sum_{j=1}^{n} \frac{(y_j-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu)^2}{2\tau^2} &= -\frac{1}{2}\sum_{j=1}^{n}\Big(\frac{(y_j-\theta)^2}{\sigma^2} + \frac{(\theta-\mu)^2}{n\tau^2}\Big) \\
&= -\frac{1}{2}\sum_{j=1}^{n}\Big(\frac{n\tau^2(y_j-\theta)^2 + \sigma^2(\theta-\mu)^2}{n\sigma^2\tau^2}\Big) \\
&= -\frac{(n\tau^2+\sigma^2)}{2\sigma^2\tau^2}\Big((n\tau^2+\sigma^2)\theta^2 - 2\theta(\mu\sigma^2 + n\bar{y}\tau^2) + n\tau^2\bar{y^2} + \sigma^2\mu^2\Big) \\
&= -\frac{(n\tau^2+\sigma^2)}{2\sigma^2\tau^2}\Big(\theta^2 - 2\theta\frac{(\mu\sigma^2 + n\bar{y}\tau^2)}{(n\tau^2+\sigma^2)} + \frac{n\tau^2\bar{y^2} + \sigma^2\mu^2}{(n\tau^2+\sigma^2)}\Big) \\
&= -\frac{\Big(\theta^2 - 2\theta\frac{(\mu\sigma^2+n\bar{y}\tau^2)}{(n\tau^2+\sigma^2)} + \Big(\frac{(\mu\sigma^2+n\bar{y}\tau^2)}{(n\tau^2+\sigma^2)}\Big)^2 - \Big(\frac{(\mu\sigma^2+n\bar{y}\tau^2)}{(n\tau^2+\sigma^2)}\Big)^2 + \frac{n\tau^2\bar{y^2}+\sigma^2\mu^2}{(n\tau^2+\sigma^2)}\Big)}{2\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}} \\
&\propto -\frac{\Big(\theta - \frac{\mu\sigma^2+n\tau^2\bar{y}}{n\tau^2+\sigma^2}\Big)^2}{2\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}}
\end{aligned}
$$

Which results in a normally distributed posterior:

$$
\theta|y \sim N(\mu_n = \frac{\frac{\mu}{\tau^2} + \frac{n\bar{y}}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \sigma_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}})
$$

Here you may notice that the forms are different. I believe the form presented here makes much more sense. It essentially states that the mean of the posterior is a convex combination between the prior and the data.

(b) A new student is sampled at random from the same population and has a weight of $\tilde{y}$ pounds. Give a posterior predictive distribution for $\tilde{y}$. (Your answer will still be a function of n.)

Solution:

We can easily find the posterior predictive distribution via integration.

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\tilde{y}-\theta)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\sigma_n^2}} e^{-\frac{(\theta-\mu_n)^2}{2\sigma_n^2}} d\theta
\end{aligned}
$$

It is easy to check for oneself, but here we are simply marginalizing over θ which will result in a gaussian for $\tilde{y}$. Therefore we just need to find the mean and variance, which can be done quite easily:

$$
\begin{aligned}
E(\tilde{y}|y) &= E_\theta(E(\tilde{y}|\theta, y)|y) \\
&= E(\theta|y) \\
&= \mu_n \\
Var(\tilde{y}|y) &= E_\theta(var(\tilde{y}|\theta, y)|y) + var(E(\tilde{y}|\theta, y)|y) \\
&= E_\theta(\sigma^2|y) + var(\theta|y) \\
&= \sigma^2 + \sigma_n^2
\end{aligned}
$$

This provides us with a predictive posterior distribution

$$\tilde{y}|y \sim N(\mu_{\tilde{y}|y} = \mu_n, \sigma^2_{\tilde{y}|y} = \sigma^2 + \sigma_n^2)$$

(c) For $n = 10$, give a 95% posterior interval for θ and a 95% posterior predictive interval for $\tilde{y}$.

Solution:

This can be done analytically since the distribution is nice. The α-level posterior predictive interval for $\theta|y$ is simply the interval $[\mu_n - z_{\frac{\alpha}{2}}\sigma_n, \mu_n - z_{1-\frac{\alpha}{2}}\sigma_n]$ while the posterior predictive interval for $\tilde{y}|y$ is $[\mu_{\tilde{y}|y} - z_{\frac{\alpha}{2}}\sigma_{\tilde{y}|y}, \mu_{\tilde{y}|y} - z_{1-\frac{\alpha}{2}}\sigma_{\tilde{y}|y}]$. $\alpha = 0.05$ will correspond to a 95% posterior predictive interval. This is most easily done in R or python.

The 95% Posterior predictive interval when $n = 10$ for $\theta|y$ is $[138.48, 162.98]$

The 95% Posterior predictive interval when $n = 10$ for $\tilde{y}|y$ is $[109.66, 191.80]$

(d) Do the same for n = 100.

Solution:

The 95% Posterior predictive interval when $n = 100$ for $\theta|y$ is $[146.16, 153.99]$

The 95% Posterior predictive interval when $n = 100$ for $\tilde{y}|y$ is $[110.68, 189.47]$

Problem 7: BDA 3rd Ed. 2.10

Discrete sample spaces: suppose there are N cable cars in San Francisco, numbered sequentially from 1 to N. You see a cable car at random; it is numbered 203. You wish to estimate N. (See Goodman, 1952, for a discussion and references to several versions of this problem, and Jeffreys, 1961, Lee, 1989, and Jaynes, 2003, for Bayesian treatments.)

(a) Assume your prior distribution on N is geometric with mean 100; that is,

$$p(N) = (1/100)(99/100)^{N-1}, \text{ for } N = 1, 2, \ldots$$

What is your posterior distribution for N?

Solution:

This problem is essentially a variation of the famous German Tank Problem. Firstly we need to establish a likelihood for this problem. We have N cable cars numbered sequentially from 1 to N, but randomly observe one to be the number 203. This tells us that there are at least 203 cable cars. More generally, we can write

$$p(X|N) = \begin{cases} \frac{1}{N} & \text{if } X \leq N \\ 0 & \text{for } X > N \end{cases}$$

The probability of observing X if $X > N$ is zero because we can not observe a label greater than the maximum value of N. If $X \leq N$, then the probability of observing X is $\frac{1}{N}$ since the trains are labeled sequentially; you have an equal chance of seeing any train number.

Now this problem doesn't admit a known distribution so we have to calculate the normalizing constant, $p(X) = \sum_{N=203}^{\infty} \frac{1}{N} \frac{1}{100} (\frac{99}{100})^{N-1}$. The easiest, and most efficient, method is to simply calculate the sum in R and Python, the code is provided below.

The calculation of the sum yields $p(X) \approx 0.00047$ which then provides us that $p(N|X) = \frac{21.25}{N} (\frac{99}{100})^{N-1}$.

(b) What are the posterior mean and standard deviation of N? (Sum the infinite series analytically or approximate them on the computer.)

Solution:

The easiest way, again, is to calculate the expected value and standard deviation in R or Python. The formula to calculate these quantities is shown below.

$$\begin{aligned} E(N|X) &= \sum_{N=203}^{\infty} N \frac{21.25}{N} (\frac{99}{100})^{N-1} \\ &\approx 279.09 \\ sd(N|X) &= \sqrt{Var(N|X)} \\ &= \sqrt{\sum_{N=203}^{\infty} (N - E(N \mid X))^2 p(N \mid X)} \\ &\approx 79.96 \end{aligned}$$

For those who are curious, though, we can actually get analytical answers to the expected values and standard deviation though I stress that it is substantially easier to compute these quantities. For simplicity, call $r = \frac{99}{100}$ and $a = \frac{1}{100p(X)}$. The expected value reduces down to a geometric series:

$$\begin{aligned} E(N|X) &= \sum_{N=203}^{\infty} ar^{N-1} \\ &= \sum_{N=1}^{\infty} ar^{N-1} - \sum_{N=1}^{202} ar^{N-1} \\ &= \frac{a}{1-r} - a\frac{1-r^{202}}{1-r} \\ &= a\frac{r^{202}}{1-r} \end{aligned}$$

We can also gain an analytic form for the standard deviation, but this will require a bit more machinery. We begin with the calculation of $E(N^2|X)$

$$\begin{aligned} E(N^2|X) &= \sum_{N=1}^{\infty} aNr^{N-1} - \sum_{N=1}^{202} aNr^{N-1} \\ &= \sum_{N=1}^{\infty} a\frac{d}{dr}r^N - \sum_{N=1}^{202} a\frac{d}{dr}r^N \\ &= \frac{d}{dr}\sum_{N=0}^{\infty} ar^N - \frac{d}{dr}\sum_{N=0}^{202} ar^N \\ &= \frac{d}{dr}\frac{a}{1-r} - \frac{d}{dr}a\frac{1-r^{203}}{1-r} \\ &= a\frac{203r^{202} - 202r^{203}}{(1-r)^2} \end{aligned}$$

Interchanging the derivative and the summation for the infinite series is justified by the uniform convergence of the geometric series. So we reach the conclusion that the standard deviation is of the form:

$$sd(N|X) = \sqrt{203a - 202ar - a^2r^{202}}\frac{r^{101}}{1-r}$$

(c) Choose a reasonable 'noninformative' prior distribution for N and give the resulting posterior distribution, mean, and standard deviation for N.

Solution:

For this problem I will discuss three potential noninformative prior distributions for N. The first choice is the simplest one, a flat prior $p(N) \propto 1$ over the natural numbers. However, one can instantly see that this leads to an improper posterior, followed by an infinite mean and variance.

$$p(X) = \sum_{N=203}^{\infty} \frac{1}{N}$$
$$E(N|X) = \sum_{N=203}^{\infty} 1$$
$$E(N^2|X) = \sum_{N=203}^{\infty} N$$

The second choice is a discrete uniform prior $p(N) \propto \frac{1}{N}$ over the natural numbers. This provides us with a proper posterior as $\sum_{N=1}^{\infty} \frac{1}{N^2} = \frac{\pi^2}{6}$ (Some of you might recognize this as the famous Basel Problem As the sum is convergent, it won't matter where we start the summation from so the posterior is proper. However, we run into the same problem as we did with our first prior choic in that our mean and variance don't exists. One should not be shocked by this. Recall distributions with proper density functions can fail to have a mean and variance. A prime example is the Cauchy distribution.

The third possible choice for an uninformative prior is substantially more interesting. We again use a flat prior, however we bound the size of N by an upper bound Ω. That is to say we define $p(N) = \frac{1}{\Omega}$ for $1 \leq N \leq \Omega$. This provides us with

$$\begin{aligned} p(X) &= \sum_{N=203}^{\Omega} \frac{1}{N} \\ &= \sum_{N=203}^{\Omega} \frac{1}{N} - \sum_{N=1}^{202} \frac{1}{N} \\ &= H_{\Omega} - H_{202} \\ &\approx log(\frac{\Omega}{202}) \end{aligned}$$

Where $H_n = \sum_{j=1}^{n} \frac{1}{j}$ is called a harmonic number. For large enough n, $H_n \approx \gamma + log(n) - \frac{1}{2n}$, where $\gamma \approx 0.577$ is the Euler-Mascheroni Constant. With this formulation the expectation and standard deviation are easy calculations.

$$\begin{aligned} E(N|X) &\approx \sum_{N=203}^{\Omega} \frac{1}{log(\frac{\Omega}{202})} \\ &\approx \frac{\Omega - 202}{log(\frac{\Omega}{202})} \\ E(N^2|X) &\approx \sum_{N=203}^{\Omega} \frac{N}{log(\frac{\Omega}{202})} \\ &\approx \frac{\frac{\Omega(\Omega+1)}{2} - \frac{202(203)}{2}}{log(\frac{\Omega}{202})} \\ sd(N|X) &\approx \sqrt{\frac{\frac{\Omega(\Omega+1)}{2} - \frac{202(203)}{2}}{log(\frac{\Omega}{202})} - \Big(\frac{\Omega - 202}{log(\frac{\Omega}{202})}\Big)^2} \end{aligned}$$

Choosing $\Omega = 500$ (there's nothing special about this number. I just thought it was reasonable), we see that

$$E(N|X) \approx 325.22$$
$$sd(N|X) \approx 86.4$$

Lastly, another reasonable choice of an uninformative prior is one that comes from a poisson distribution, $p(N) \propto \frac{(100)^N e^{-100}}{N!}$. The posterior for this choice in prior is proper and yields the following mean and standard deviation:

$$E(N|X) \approx 203.93$$
$$sd(N|X) \approx 1.34$$

The mean is too close to that of the data, with such a small standard deviation.

Note: You'll notice that the poisson pmf involves a factorial which blows up for large values, and thus is hard to calculate in R and python. One should use lfactorial in R, and gammaln in Python when computing the log of factorials.

Problem 8: BDA 3rd Ed. 2.13

Discrete data: The table below gives the number of fatal accidents and deaths on scheduled airline flights per year over a ten-year period. We use these data as a numerical example for fitting discrete data models.

Table 1. Worldwide airline fatalities, 1976–1985. Death rate is passenger deaths per 100 million passenger miles. Source: Statistical Abstract of the United States.

Years	Fatal accidents	Passenger deaths	Death rates
1976	24	734	0.19
1977	25	516	0.12
1978	31	754	0.15
1979	31	877	0.16
1980	22	814	0.14
1981	21	362	0.06
1982	26	764	0.13
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

(a) Assume that the numbers of fatal accidents in each year are independent with a Poisson(θ) distribution. Set a prior distribution for θ and determine the posterior distribution based on the data from 1976 through 1985. Under this model, give a 95% predictive interval for the number of fatal accidents in 1986. You can use the normal approximation to the gamma and Poisson or compute using simulation.

Solution:

Let $y_j|\theta \sim Poisson(\theta)$ where $j = 1, 2, \cdots, 10$ is the corresponding year (i.e. $j = 1$ refers to 1976). We will simplify our calculations by using the conjugate prior $\theta \sim Gamma(\alpha, \beta)$ which will then yield the posterior distribution $\theta|y \sim Gamma(\sum_{j=1}^{n} y_j + \alpha, n + \beta)$ where n is our

number of observations. Using a non-informative prior by setting $\alpha = \beta = 0$. We will do a quick derivation to show that this is a conjugate prior.

$$\begin{aligned} p(\theta|y) &\propto \Big(\prod_{j=1}^{n} \theta^{y_j} e^{-\theta}\Big)\theta^{\alpha-1}e^{-\beta\theta} \\ &\propto \theta^{\sum_{j=1}^{n} y_j+\alpha-1}e^{-(n+\beta)\theta} \end{aligned}$$

Which is a Gamma kernel so we conclude that $\theta|y \sim Gamma(\sum_{j=1}^{n} y_j + \alpha, n + \beta)$. For the rest of this problem and the other parts we will use an improper prior by setting $\alpha = \beta = 0$. This gives us a posterior distribution $\theta|y \sim Gamma(238, 10)$.

Disclaimer: This results in $\theta \sim Gamma(0, 0)$. This is simply a placeholder for the improper prior $p(\theta) \propto \frac{1}{\theta}$ whose kernel resembles that of a Gamma Distributions. I stress that it is not an actual Gamma distribution; it's simply just useful notation. In fact, Gamma(0,0) makes no sense as a distribution.

Now we wish to gain a posterior predictive interval for $\tilde{y}|y$. Luckily we have two nice, known distributions whose posterior predictive distribution is a well known distribution as well. We will prove it right now. For generality, we will use this form for the likelihood distribution $\tilde{y} \mid \theta \sim Poisson(m\theta)$. You will see why this will be helpful in parts (b) and (d).

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &= \int \frac{(m\theta)^{\tilde{y}}e^{-m\theta}}{\tilde{y}!}\frac{(n+\beta)^{n\bar{y}+\alpha}}{\Gamma(n\bar{y}+\alpha)}\theta^{n\bar{y}+\alpha-1}e^{-(n+\beta)\theta}d\theta \\ &= \frac{m^{\tilde{y}}(n+\beta)^{n\bar{y}+\alpha}}{\tilde{y}!\Gamma(n\bar{y}+\alpha)}\int \theta^{n\bar{y}+\alpha+\tilde{y}-1}e^{-(n+\beta+m)\theta}d\theta \\ &= \frac{m^{\tilde{y}}(n+\beta)^{n\bar{y}+\alpha}}{\tilde{y}!\Gamma(n\bar{y}+\alpha)}\frac{(n+\beta+m)^{n\bar{y}+\alpha+\tilde{y}}\Gamma(n\bar{y}+\alpha+\tilde{y})}{(n+\beta+m)^{n\bar{y}+\alpha+\tilde{y}}\Gamma(n\bar{y}+\alpha+\tilde{y})}\int \theta^{n\bar{y}+\alpha+\tilde{y}-1}e^{-(n+\beta+m)\theta}d\theta \\ &= \frac{\Gamma(n\bar{y}+\alpha+\tilde{y})}{\tilde{y}!\Gamma(n\bar{y}+\alpha)}\Big(\frac{n+\beta}{n+\beta+m}\Big)^{n\bar{y}+\alpha}\Big(\frac{m}{n+\beta+m}\Big)^{\tilde{y}} \\ &\sim NB(r = n\bar{y}+\alpha, p = \frac{n+\beta}{n+\beta+m}) \end{aligned}$$

This reduces to a negative binomial by noticing the kernel in the integral is that of a gamma with shape=$n\bar{y} + \alpha + \tilde{y}$ and rate = $n + \beta + m$. Using the negative binomial distribution, with $m = 1$, we see that a 95% predictive interval for $\tilde{y}|y$ is [14, 34].

Note: You will not always be able to get such a nice form for the posterior predictive distribution. A common way to draw samples from the distribution $\tilde{y}|y$ is to draw samples from the posterior $\theta|y$, and plug these samples back into the likelihood distribution $y|\theta$ in order to draw samples $\tilde{y}$. Both methods are shown in R and Python in the appendix, and the output applies to parts (a)-(d).

(b) Assume that the numbers of fatal accidents in each year follow independent Poisson distributions with a constant rate and an exposure in each year proportional to the number of passenger miles flown. Set a prior distribution for θ and determine the posterior distribution based on the data for 1976–1985. (Estimate the number of passenger miles flown in each year by dividing the appropriate columns of Table 2.2 and ignoring round-off errors.) Give a 95% predictive interval for the number of fatal accidents in 1986 under the assumption that 8×10^{11} passenger miles are flown that year.

Solution:

We obtain the number of passenger miles flown in each year by dividing the "passenger deaths" column by the "death rate" column, and then multiply by 100 million, since death rate is measured per 100 million passenger miles. The results are given in the table below:

Years	Estimated Passenger Miles
1976	3.9×10^{11}
1977	4.3×10^{11}
1978	5.0×10^{11}
1979	5.5×10^{11}
1980	5.8×10^{11}
1981	6.0×10^{11}
1982	5.9×10^{11}
1983	6.2×10^{11}
1984	7.4×10^{11}
1985	7.1×10^{11}

Let m_j denote the estimated number of miles flown in year j. $y_j|m_j\theta \sim Poisson(m_j\theta)$ where $j = 1, 2, \cdots, 10$. We will again use the non-informative prior $\theta \sim Gamma(0, 0)$.

$$\begin{aligned} p(\theta|y) &\propto \Big(\prod_{j=1}^{n} p(y_j|\theta)\Big)p(\theta) \\ &= \Big(\prod_{j=1}^{n} \frac{(m_j\theta)^{y_j}e^{m_j\theta}}{y_j!}\Big)\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ &\propto \theta^{\sum_{j=1}^n y_j+\alpha-1}e^{-(\beta+\sum_{j=1}^n m_j)\theta} \\ &\sim Gamma(\sum_{j=1}^{n} y_j + \alpha, \beta + \sum_{j=1}^{n} m_j) \end{aligned}$$

Our posterior distribution is $\theta|y \sim Gamma(238, 5.7 \times 10^{12})$. Following the derivation of the posterior predictive distribution from part (a), we see that $\tilde{y}|y \sim NB(r = n\bar{y} + \alpha, p = \frac{\sum_{j=1}^{10} m_j+\beta}{\sum_{j=1}^{10} m_j+\beta+m})$, where m corresponds to the miles flown in 1986. Using the negative binomial distribution, we see that a 95% predictive interval for $\tilde{y}|y$ is [22, 46].

Note: Another valid way to do this is to simply draw samples from the posterior, and use these samples to draw from the likelihood for 1986.

(c) Repeat (a) above, replacing 'fatal accidents' with 'passenger deaths.'

Solution:

Simply replace 238 with 6919 in the Gamma Posterior and the Negative Binomial Predictive Posterior. Our posterior distribution is $\theta|y \sim Gamma(6919, 10)$. Following the derivation of the posterior predictive distribution from part (a), we see that $\tilde{y}|y \sim NB(r = n\bar{y} + \alpha, p = \frac{n+\beta}{n+\beta+1})$. Using the negative binomial distribution, we see that a 95% predictive interval for $\tilde{y}|y$ is [639, 747].

(d) Repeat (b) above, replacing 'fatal accidents' with 'passenger deaths.'

Solution:

Again, replace 238 with 6919 in the Gamma Posterior and the Negative Binomial Predictive Posterior. Our Posterior Distribution is $\theta|y \sim Gamma(6919, 5.7\times10^{12})$. Following the derivation of the posterior predictive distribution from part (a), we see that $\tilde{y}|y \sim NB(r = n\bar{y} + \alpha, p = \frac{\sum_{j=1}^{10} m_j+\beta}{\sum_{j=1}^{10} m_j+\beta+m_{11}})$, where m corresponds to the miles flown in 1986. Using the negative binomial distribution, we see that a 95% predictive interval for $\tilde{y}|y$ is [906, 1037].

(e) In which of the cases (a)–(d) above does the Poisson model seem more or less reasonable? Why? Discuss based on general principles, without specific reference to the numbers in Table 2.2. Incidentally, in 1986, there were 22 fatal accidents, 546 passenger deaths, and a death rate of 0.06 per 100 million miles flown. We return to this example in Exercises 3.12, 6.2, 6.3, and 8.14.

Solution:

There are a number of qualities which make each model reasonable and unreasonable. Let's begin by comparing the Poisson models for Fatal Accidents (a) and Passenger Deaths (c).

A fatal accident on a plane could be due to any number of reasons. Perhaps there were mechanical errors, or maybe errors made by the pilots, or even extreme weather conditions could be at fault. The point is that one fatal accident is more likely than not to be independent of another fatal accident. One won't cause the other. On the other hand, Passenger deaths aren't necessarily independent. They occur in clusters so there isn't any spacing between events like there could be in fatal accidents; everyone dies at the same during the crash so it's misleading to model this using a Poisson distribution where we have a rate parameter. Perhaps the Poisson Distribution is better suited for modeling fatal accidents than it is for passenger deaths.

Next let's compare the reasonableness of modeling using a constant rate vs an exposure proportional to the number of passenger miles flown. The latter is makes more intuitive sense in both cases; the more you fly the more likely it is you are involved in a fatal accident. Likewise the more you fly the more you run the risk of involving yourself in more fatal accidents, and thereby increasing the number of passenger deaths.

Overall, to make a long story short, models (b) and (d) make more sense over (a) and (c) as they take into account the number of miles flown. However, (a) and (b) are more reasonable over (c) and (d) since fatal accidents are more likely to fit a Poisson model over passenger deaths.

Problem 9: BDA 3rd Ed. 2.17

Posterior intervals: unlike the central posterior interval, the highest posterior interval is not invariant to transformation. For example, suppose that, given σ^2, the quantity $\frac{n\nu}{\sigma^2}$ is distributed as χ^2_n, and that σ has the (improper) noninformative prior density $p(\sigma) \propto \sigma^{-1}$, $\sigma > 0$.

(a) Prove that the corresponding prior density for σ^2 is $\mathrm{p}(\sigma^2) \propto \sigma^{-2}$.

Solution:

This is just a simple change of variables. Let $\theta = \sigma^2$ which yields $\frac{d\sigma}{d\theta} = \frac{1}{2\sqrt{\theta}}$.

$$\begin{aligned} p(\theta) = p(\sigma^2) &= \left|\frac{d\sigma}{d\theta}\right| p(\sigma(\theta)) \\ &\propto \frac{1}{2\sqrt{\theta}}\frac{1}{\sqrt{\theta}} \\ &\propto \frac{1}{\sigma^2} \end{aligned}$$

(b) Show that the 95% highest posterior density region for σ^2 is not the same as the region obtained by squaring the endpoints of a posterior interval for σ.

Solution:

Let's assume the result is true, and find a contradiction. We will begin by finding the posterior density functions. Recall that a χ^2_n distribution is equivalent to a $Gamma(\frac{n}{2}, 2)$ which is a scale family, i.e. $g(x \mid \theta) = \frac{1}{\theta} f(\frac{x}{\theta})$, where g is the pdf of $x \mid \theta$, and f is the probability density function associated with $\frac{x}{\theta}$. In this case, let $x = n\nu$, and let $\theta = \sigma^2$

We're aware that $\frac{n\nu}{\sigma^2} \sim \chi^2_n$. This gives us

$$\begin{aligned} p(\frac{n\nu}{\sigma^2}) &= \frac{(\frac{n\nu}{\sigma^2})^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} e^{-\frac{\frac{n\nu}{\sigma^2}}{2}} \\ &\propto (\frac{n\nu}{\sigma^2})^{\frac{n}{2}-1} e^{-\frac{\frac{n\nu}{\sigma^2}}{2}} \end{aligned}$$

Using the scale family equation we can see that the likelihood for this function takes the form

$$\begin{aligned} f(n\nu \mid \sigma^2) &= \frac{1}{\sigma^2} p(\frac{n\nu}{\sigma^2}) \\ &= \frac{1}{\sigma^2} (\frac{n\nu}{\sigma^2})^{\frac{n}{2}-1} e^{-\frac{n\nu}{2\sigma^2}} \end{aligned}$$

Now to find our posterior distributions $p(\sigma \mid n\nu)$ and $p(\sigma^2 \mid n\nu)$, we proceed as follows:

$$\begin{aligned} p(\sigma^2 \mid n\nu) &\propto p(n\nu \mid \sigma^2) p(\sigma^2) \\ &= \frac{1}{\sigma^2} (\frac{n\nu}{\sigma^2})^{\frac{n}{2}-1} e^{-\frac{n\nu}{2\sigma^2}} \frac{1}{\sigma^2} \\ &\propto (\frac{1}{\sigma^2})^2 (\frac{1}{\sigma^2})^{\frac{n}{2}-1} e^{-\frac{n\nu}{2\sigma^2}} \\ &= (\frac{1}{\sigma^2})^{\frac{n}{2}+1} e^{-\frac{n\nu}{2\sigma^2}} \\ &= (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{n\nu}{2\sigma^2}} \end{aligned}$$

A change of variables via $u = \sigma$ will yield

$$\begin{aligned} p(\sigma \mid n\nu) &\propto \left| \frac{d\sigma^2}{du} \right| p(\sigma^2 \mid n\nu) \\ &= \sigma (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{n\nu}{2\sigma^2}} \\ &\propto \sigma^{-n-1} e^{-\frac{n\nu}{2\sigma^2}} \end{aligned}$$

Now suppose that the interval (a,b) is the 95% HPD region for $\sigma^2 | n\nu$. Assuming the result is true, then $(\sqrt{a}, \sqrt{b})$ is the 95% HPD region for $\sigma | n\nu$.

The posterior density at the endpoints of the HPD are equal which implies

$$\begin{aligned} (a)^{-\frac{n}{2}-1} e^{-\frac{n\nu}{2(a)}} &= (b)^{-\frac{n}{2}-1} e^{-\frac{n\nu}{2(b)}} \\ \sqrt{a}^{-n-1} e^{-\frac{n\nu}{2a}} &= \sqrt{b}^{-n-1} e^{-\frac{n\nu}{2b}} \end{aligned}$$

Following a bit of algebra we see that

$$-(\frac{n}{2}+1)log(a)-\frac{n\nu}{2a}=-(\frac{n}{2}+1)log(b)-\frac{n\nu}{2b}$$
$$-(\frac{n+1}{2})log(a)-\frac{n\nu}{2a}=-(\frac{n+1}{2})log(b)-\frac{n\nu}{2b}$$

Solving this system of equations results in $a = b$, a null interval, which is a contradiction.

Problem 10: BDA 3rd Ed. 2.20

Censored and uncensored data in the exponential model:

(a) Suppose $y|\theta$ is exponentially distributed with rate θ, and the marginal (prior) distribution of θ is Gamma(α, β). Suppose we observe that $y \geq 100$, but do not observe the exact value of y. What is the posterior distribution, $p(\theta|y \geq 100)$, as a function of α and β? Write down the posterior mean and variance of θ.

Solution:

This is a survival analysis problem, and I understand that some of you probably have never taken a survival course. For that reason I will very briefly introduce the logic used to set up the likelihood.

Survival analysis deals with time to event data. A critical function is the survival function $S(t) = P(X > t)$, the probability someone survives past time t. Survival Analysis also deals with censored data which is simply data where you don't observe the event of interest due to a person dropping out of the study (e.g. Following a person until they develop lung cancer for 5 years. If they develop cancer after year 5, you never observe it). The observations we deal with are joint probabilities, i.e. $Pr(y, \delta \mid \theta) = p(y \mid \theta)^{\delta} S(y \mid \theta)^{1-\delta}$, where δ is the censoring mechanism, 1 if we see they experience the event, 0 otherwise. This says that if we observe a subject at time y, they're contribution to the likelihood is the probability density function $p(y \mid \theta)$. If the subject still hasn't experienced the event at y, all we know under non-informative censoring is that their lifetime exceeds y, so the likelihood becomes $S(y \mid \theta) = P(T > y \mid \theta)$.

We don't observe y, only that y exceeds or equals 100. Our likelihood reduces to the survival function. Recall for an exponential distribution that $P(T \leq y|\theta) = 1 - e^{-\theta y}$. Using this we can see that our posterior density will be of the form:

$$\begin{aligned} p(\theta|y \geq 100) &\propto P(y \geq 100|\theta)p(\theta|\alpha, \beta) \\ &\propto e^{-100\theta}\theta^{\alpha-1}e^{-\beta\theta} \end{aligned}$$

Which gives us that $\theta|y \geq 100 \sim Gamma(\alpha, \beta + 100)$. Knowing the exact distribution makes this a swift calculation of the mean and variance:

$$E(\theta|y \geq 100) = \frac{\alpha}{\beta + 100}$$
$$Var(\theta|y \geq 100) = \frac{\alpha}{(\beta + 100)^2}$$

(b) In the above problem, suppose that we are now told that y is exactly 100. Now what are the posterior mean and variance of θ?

Solution:

Now that we observe y we use the probability density function which is of the form $p(y|\theta) = \theta e^{-\theta y}$. Following the same procedure we see that $\theta|y = 100 \sim Gamma(\alpha + 1, \beta + 100)$. The mean and variance are presented below:

$$E(\theta|y = 100) = \frac{\alpha + 1}{\beta + 100}$$
$$Var(\theta|y = 100) = \frac{\alpha + 1}{(\beta + 100)^2}$$

(c) Explain why the posterior variance of θ is higher in part (b) even though more information has been observed. Why does this not contradict identity (2.8) on page 32?

Solution:

The essence of this problem is that knowing $y = 100$ provides substantially more information than if we know that $y \geq 100$. It's akin to asking someone where New Orleans is located. $y \geq 100$ is someone telling you it's in the Southern part of the United States. $y = 100$ is akin to someone telling you it's located in Louisiana, a U.S. state, right on the Mississippi River.

Thus we're left befuddled as the variance of the $y = 100$ model is greater than the variance of the $y \geq 100$ model. The more informative model gives us exactness so we would expect a very small variance, while the less informative provides us with a large range to check over, and we would expect more variance.

Recall that identity 2.8 refers to the following formula:

$$Var(\theta) = E(Var(\theta|y)) + Var(E(\theta|y))$$
$$Var(\theta) \geq E(Var(\theta|y))$$

This essentially states that if we average the variance over all possible posterior models (note the expected value on the right is taken with respect to y), the resulting average variance will be less than the variance of the prior distribution.

This, however, doesn't really get us anywhere so let's employ the law of total conditional variance to see why . The law of Total Conditional variance states. (Note: the variable in the conditional is just shorthand for the sigma-algebra).

$$Var(\theta|X_1) = E(Var(\theta|X_1, X_2)|X_2) + Var(E(\theta|X_1, X_2)|X_2)$$

Allow $X_1 = y$ and $X_2 = y \geq k$ for $k \in [0, \infty)$. Since $\sigma(X_2) \subset \sigma(X_1)$, the law of total conditional variance simplifies to the below model.

$$Var(\theta \mid y \geq k) \geq E(Var(\theta|y)|y \geq k)$$

Now note the expected value on the right is taken with respect to the posterior predictive distribution $\tilde{y} \mid y \geq k$. The left hand side is the variance of a posterior model where we know $y \geq k$. The right hand side is the average of the posterior variance over all possible models resulting from $y \geq k$. When $k = 100$ our inequality becomes

$$Var(\theta \mid y \geq 100) \geq E(Var(\theta|y)|y \geq 100)$$

Which says the variance of the model in (a) is greater than the average of the posterior variance over all models where $y \geq 100$. There is no contradiction to be had here; One might think that we can plug $y = 100$ into the right side, and get that $Var(\theta \mid y \geq 100) \geq E(Var(\theta|y = 100)|y \geq 100) = Var(\theta|y = 100)$ but that is false. Plugging in $y = 100$ is not the same as averaging the posterior variance over the distribution $\tilde{y} \mid y \geq 100$, which is what the right side of the inequality does.