

Instructions

- (1) Homework must be typed and answered in the order given (problem 1(a)(b)(c)(d) first, problem 2(a)(b)... second, etc...)
- (2) Undergrads and grads will answer all questions.
- (3) Include in each part of the homework only the answer. R code and R output (without mistakes), must be included in the appendix to the question. For example, for question 1.a, write only the answer and your comments. The code and output for that part of the question will be in the appendix (the last part of question 1).
- (4) No late homework under any circumstances.
- (5) Write your name and ID this way: Last name, first name, UCLA ID, date, Homework number.
- (6) Do not just give a number as an answer. For example, if asked for probability that posterior proportion is larger than 0.7, write $Prob(p > 0.7) = 0.3$, say and write comments or explanations if needed.
- (7) The homework must be turned in in lecture (no mail box, no e-mail).

Problem 1. Use material from Outline 5, section 7.

Suppose the causes of death are reviewed in detail for a city in the United States for a single year. It is found that 10 persons, out of a population of 500,000, died of skin cancer, giving an estimated cancer mortality rate in the city of 2 cases per 100,000 persons per year. A Poisson sampling model is often used for epidemiological data of this form. The Poisson model derives from an assumption of exchangeability among all small intervals of exposure. Define y as the number of deaths in a city of 500,000 in one year, and θ as the true underlying long-term cancer mortality rate in the city (measured in cases per 100,000 persons per year). The model for the data is then Poisson(5θ).

We can use knowledge about skin cancer mortality rates around the world to construct a prior distribution for θ and then combine the datum $y = 10$ with that prior distribution to obtain a posterior distribution.

- (a) What is a sensible Gamma prior distribution for θ ? Google a little to find mortality rates around the world, and potential information about probabilities of rates being larger or smaller than 2 cases per thousand, probabilities of being larger than 3, etc.... Enough information to determine whether there is a right tail or left tail, or whether the distribution is symmetric. Give the (a, b) of your gamma prior, the mean, and some probabilities.
- (b) What is the posterior distribution for θ (specify posterior parameters of the posterior Gamma distribution).
- (c) Compute the posterior probability that the long-term death rate from skin cancer in our city is more than 1.0 per 100,000 per year.

Solution 1. (a) $y = 10$ is a single observation with exposure $x = 5$

According to the World Health Organization (WHO),

http://en.wikipedia.org/wiki/Skin_cancer#cite_note-24

and

http://www.who.int/healthinfo/global_burden_disease/estimates_country/en/index.html

more than 50% of the world have rates lower than 4 cancers per 100,000 people. There are some cancer rates up to 7.7. The mean is probably around 3 cases per 100000 with a standard deviation of about 1.5. I decide that a Gamma(6,2) prior might be appropriate. This has the following summaries:

Mean= $6/2=3$; Variance= $6/4=1.5$; $sd=1.22$; $Prob(rate < 4) = 0.80876$

(b) Gamma($6+10, 2+5$) = Gamma(16,7)

(c) $Prob(rate > 1 \text{ per } 10000) = 0.99759$

Posterior mean is $16/7 = 2.2857$

Problem 2. The table below gives the number of fatal accidents and deaths on scheduled airline flights per year over a ten-year period.

- Assume that the number 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 passenger deaths in 1986.
- 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 the table and ignoring round-off errors.) Give a 95% predictive interval for the number of passenger deaths in 1986 under the assumption that 8000 million passenger miles are flown that year.
- Repeat (a) above, replacing "fatal accidents" with "passenger deaths."
- Repeat (b) above, replacing "fatal accidents" with "passenger deaths."

Year	Fatal accidents	Passenger deaths	Death rate
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	0.33
1983	20	809	0.13
1984	16	223	0.03
1985	22	1066	0.15

Problem 3. In section 5.9, we conclude the problem with the following statements:

Notice that there is much more overlap between these two distributions than between the posterior distributions of θ_1 and θ_2 . For example, $\Pr(\tilde{Y}_1 > \tilde{Y}_2 \mid \sum Y_{i,1} = 217, \sum Y_{i,2}) = 0.22$. The distinction between the events $\{\theta_1 > \theta_2\}$ and $\tilde{Y}_1 > \tilde{Y}_2$ is extremely important: Strong evidence of a difference between two populations does not mean that the difference itself is large.

Use R to check whether those statements are true. The way to do that is to first generate observations from the distribution of Y_1 , then generate observations from the distribution of Y_2 and after that, count the proportion of one larger than the other. It is done in the textbook, so you may use the same commands.

Problem 4. 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.

- Give your posterior distribution for θ . (Your answer will be a function of n .)
- 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 .)
- For $n=10$, give a 95% posterior interval for θ and a 95% posterior predictive interval for \tilde{y} .

Problem 5. Computing with a nonconjugate single-parameter model. suppose that y_1, y_2, \dots, y_n are independent samples from a Cauchy distribution with unknown center θ and known scale 1.:

$$p(y_i \mid \theta) \propto 1/(1 + (y_i - \theta)^2)$$

. Assume, for simplicity, that the prior distribution for θ is uniform on $[0, 1]$. Given the observations $(y_1, y_2, \dots, y_5) = (-2, -1, 0, 1.5, 2.5)$:

(a) Compute the unnormalized posterior density function $p(\theta)p(y | \theta)$, on a grid of points $\theta = 0, 1/m, 2/m, \dots, 1$, for some large integer m . Using the grid approximation, compute and plot the normalized posterior density function, $p(\theta | y)$ as a function of θ .

(b) Sample 1000 draws of θ from the posterior density and plot a histogram of the draws.

(c) Use the 1000 samples of θ to obtain 1000 samples from the predictive distribution of a future observation, y_6 , and plot a histogram of the predictive draws.

Will this program do the job?

```
f = function(theta) { # evaluates the likelihood
  likelihood= (1/(1+(-2-theta)^2))*(1/(1+(-1-theta)^2))*(1/(1+(0-theta)^2)) *(1/(1+(1.5-theta)^2))*(1/(1+(2.5-theta)^2))
}
```

```
grid = seq(from=0, to=1, length=100) # make a grid of theta values -step=0.01
fgrid=rep(0,100) # initialize a vector for the function evaluations
```

```
for(i in 1:100){ # evaluate f on the grid
fgrid[i] = f(grid[i])
}
```

```
c=0.01*sum(fgrid) #compute the normalization constant
# dens norm = dens.unn/step x (sum(dens unnorma))
px=fgrid/c
```

```
par(mfrow=c(2,1))
plot(grid, fgrid, xlab="theta",ylab="P(theta|y)", type="l",
main="un-normalized posterior",ylim=c(0,max(fgrid)), xlim=c(0,1) )
```

```
plot(grid, px, xlab="theta",ylab="P(theta|y)", type="l",
main="normalized posterior",ylim=c(0,max(px)), xlim=c(0,1) ) #plot it
```

Solution 2.

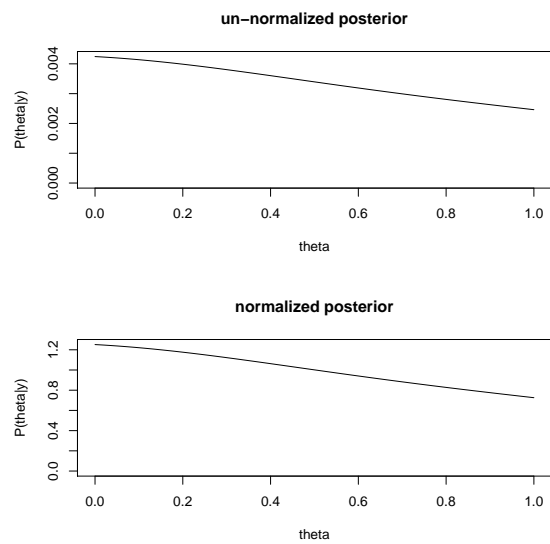


Figure 1: Unnormalized and normalized posterior distributions for a non-conjugate single-parameter model: Cauchy sampling model and $U(0,1)$ prior.