Lab 3

PSTAT 115, Winter 2022

Jan 18, 2022

This lab will focus on the followng topics:

- Derive posterior distribution from prior and likelihood function
- Common conjugate priors and the meaning of them.
- Posterior mean and MAP estimates

The following example will be adopted throughout to illustrate the concepts.

Estimating the probability of a female birth

The proportion of births that are female has long been a topic of interest both scientifically and to the lay public. Two hundred years ago it was established that the proportion of female births in European populations was less than 0.5. The currently accepted value of the proportion of female births in large European-race populations is 0.485.

- Assuming Bernoulli distribution for each child birth, and let p denote the probability of being a female. What is the probability of observing y female babies among n newborn babies (y is between 0 and n, inclusive)?

Solution: It is the binomial density

$$\binom{n}{y} p^y (1-p)^{n-y}$$

- Now assume we do not have any prior information about the probability of female birth, so a uniform prior on $[0, 1]$ is used. What is the posterior distribution under this prior?

Solution:

$$\begin{aligned} p(\theta|y) &\propto p(\theta) * p(y|\theta) \\ &= \binom{n}{y} p^y (1-p)^{n-y} \text{ (in this context } \theta \text{ is } p) \\ &\propto p^y (1-p)^{n-y} \end{aligned}$$

The above corresponds to the functional form of a beta distribution. Specifically, it is $Beta(y+1, n-y+1)$.

- We know the uniform distribution is a special case of the Beta distribution. Now what if we apply a $Beta(2,2)$ distribution as our prior? What is the posterior distribution now?

Solution:

$$\begin{aligned} p(\theta|y) &\propto p(\theta) * p(y|\theta) \\ &= p^{2-1} * (1-p)^{2-1} \binom{n}{y} p^y (1-p)^{n-y} \text{ (in this context } \theta \text{ is } p) \\ &\propto p^{y+1}(1-p)^{1+n-y} \end{aligned}$$

The above corresponds to the functional form of a beta distribution. Specifically, it is $Beta(y+2, n-y+2)$.

Common Conjugate priors

Conjugacy is formally defined as follows. A collection of pdfs (or pmfs) is called a conjugate prior family for a model $X \sim f(x|\theta), \theta \in \Theta$, if whenever a prior $\xi(\theta)$ is chosen from the collection, it leads to a posterior $\xi(\theta|x)$ that is also a member of the collection, for every observation $X = x$. For instance, in our above binomial model, the Beta distribution is a conjugate prior to the binomial distribution.

Conjugate priors permit fast posterior computations, which can be valuable in high dimension problems. Meanwhile, there are handy formulas for mean and mode of the posterior distributions. The following table summarizes the common distribution we might encounter throughout the course. The mean and variance of each posterior distribution are also of importance.

Model	Parameter	Prior	Posterior
$X \sim \text{Binomial}(n,p)$	$0 \le p \le 1$	$\text{Beta}(a,b)$ $a>0, b>0$	$\text{Beta}(a',b')$ $a' = a + x$ $b' = b + n - x$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Poisson}(\lambda)$	$\lambda > 0$	$\text{Gamma}(a,b)$ $a>0, b>0$	$\text{Gamma}(a',b')$ $a' = a + n\bar{x}$ $b' = b + n$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Exponential}(\lambda)$	$\lambda > 0$	$\text{Gamma}(a,b)$ $a>0, b>0$	$\text{Gamma}(a',b')$ $a' = a + n$ $b' = b + n\bar{x}$
$X = (X_1, \cdots, X_n)$ $X_i \overset{\text{IID}}{\sim} \text{Normal}(\mu, \sigma^2)$ σ^2 known	$-\infty < \mu < \infty$	$\text{Normal}(a, b^2)$ $-\infty < a < \infty$ $b > 0$	$\text{Normal}(a', b'^2)$ $a' = \frac{nb^2\bar{x} + \sigma^2 a}{nb^2 + \sigma^2}$ $b'^2 = \frac{\sigma^2 b^2}{nb^2 + \sigma^2}$

Continuous Distributions

distribution	pdf	mean	variance
Beta(α,β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1};\ x\in(0,1),\ \alpha,\beta>0$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Cauchy(θ,σ)	$\frac{1}{\pi\sigma}\frac{1}{1+(\frac{x-\theta}{\sigma})^2};\ \sigma>0$	does not exist	does not exist
Notes: Special case of Students's t with 1 degree of freedom. Also, if X,Y are iid $N(0,1)$, $\frac{X}{Y}$ is Cauchy			
χ^2_p	$\frac{1}{\Gamma(\frac{p}{2})2^{\frac{p}{2}}}x^{\frac{p}{2}-1}e^{-\frac{x}{2}};\ x>0,\ p\in N$	p	$2p$
Notes: Gamma$(\frac{p}{2},2)$.			
Double Exponential(μ,σ)	$\frac{1}{2\sigma}e^{-\frac{\lvert x-\mu\rvert}{\sigma}};\ \sigma>0$	μ	$2\sigma^2$
Exponential(θ)	$\frac{1}{\theta}e^{-\frac{x}{\theta}};\ x\geq 0,\ \theta>0$	θ	θ^2
Notes: Gamma$(1,\theta)$. Memoryless. $Y=X^{\frac{1}{\gamma}}$ is *Weibull*. $Y=\sqrt{\frac{2X}{\beta}}$ is *Rayleigh*. $Y=\alpha-\gamma\log\frac{X}{\beta}$ is *Gumbel*.			
F_{ν_1,ν_2}	$\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}\frac{x^{\frac{\nu_1-2}{2}}}{\left(1+(\frac{\nu_1}{\nu_2})x\right)^{\frac{\nu_1+\nu_2}{2}}};\ x>0$	$\frac{\nu_2}{\nu_2-2},\ \nu_2>2$	$2(\frac{\nu_2}{\nu_2-2})^2\frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)},\ \nu_2>$
Notes: $F_{\nu_1,\nu_2}=\frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2}$, where the χ^2s are independent. $F_{1,\nu}=t^2_\nu$.			
Gamma(α,β)	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}};\ x>0,\ \alpha,\beta>0$	$\alpha\beta$	$\alpha\beta^2$
Notes: Some special cases are exponential $(\alpha=1)$ and χ^2 $(\alpha=\frac{p}{2},\beta=2)$. If $\alpha=\frac{2}{3}$, $Y=\sqrt{\frac{X}{\beta}}$ is *Maxwell*. Y			
Logistic(μ,β)	$\frac{1}{\beta}\frac{e^{-\frac{x-\mu}{\beta}}}{\left[1+e^{-\frac{x-\mu}{\beta}}\right]^2};\ \beta>0$	μ	$\frac{\pi^2\beta^2}{3}$
Notes: The cdf is $F(x\lvert\mu,\beta)=\frac{1}{1+e^{-\frac{x-\mu}{\beta}}}$.			
Lognormal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}};\ x>0,\sigma>0$	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2(\mu+\sigma^2)}-e^{2\mu+\sigma^2}$
Normal(μ,σ^2)	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}};\ \sigma>0$	μ	σ^2
Pareto(α,β)	$\frac{\beta\alpha^\beta}{x^{\beta+1}};\ x>\alpha,\ \alpha,\beta>0$	$\frac{\beta\alpha}{\beta-1},\ \beta>1$	$\frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)},\ \beta>2$
t_ν	$\frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})}\frac{1}{\sqrt{\nu\pi}}\frac{1}{(1+\frac{x^2}{\nu})^{\frac{\nu+1}{2}}}$	$0,\ \nu>1$	$\frac{\nu}{\nu-2},\nu>2$
Notes: $t^2_\nu=F_{1,\nu}$.			
Uniform(a,b)	$\frac{1}{b-a},\ a\leq x\leq b$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
Notes: If $a=0,\ b=1$, this is special case of beta $(\alpha=\beta=1)$.			
Weibull(γ,β)	$\frac{\gamma}{\beta}x^{\gamma-1}e^{-\frac{x^\gamma}{\beta}};\ x>0,\ \gamma,\beta>0$	$\beta^{\frac{1}{\gamma}}\Gamma(1+\frac{1}{\gamma})$	$\beta^{\frac{2}{\gamma}}\left[\Gamma(1+\frac{2}{\gamma})-\Gamma^2(1+\right.$
Notes: The mgf only exists for $\gamma\geq 1$.			

Solution: There is no problem for this part. But you are encouraged to derive the conjugate pairs listed in the table to appreciate this beautiful idea.

#Meaning of Priors

- Binomial with beta prior

Pseudo-Counts Interpretation

successes ("made shots")

- Observe y successes, $n - y$ failures
- If $p(\theta) \sim \text{Beta}(\alpha, \beta)$ then $p(\theta \mid y) = \text{Beta}(y + \alpha, n - y + \beta)$

units should be same

failures "missed shots"

- What is $E[\theta \mid y]$?

α: "prior successes" "pseudo-counts" of success

β: "prior failures"

$$\frac{\alpha}{\alpha+\beta} = \frac{\text{successes}}{\text{successes + failures}}$$

$$P(\theta|y) \sim \text{Beta}(y+\alpha, n-y+\beta)$$

$$E[\theta|y] = \frac{y+\alpha}{y+\alpha+n-y+\beta} = \frac{y+\alpha}{n+\alpha+\beta}$$

$$= \frac{n}{n}\frac{y}{n+\alpha+\beta} + \frac{\alpha}{n+\alpha+\beta}\frac{\alpha+\beta}{\alpha+\beta}$$

$$= \frac{n}{n+\alpha+\beta} \times \frac{y}{n} + \frac{\alpha+\beta}{n+\alpha+\beta} \times \frac{\alpha}{\alpha+\beta}$$

$$= \frac{n}{n+\alpha+\beta}\hat{\theta}_{MLE} + \frac{\alpha+\beta}{n+\alpha+\beta}\hat{\theta}_{\text{prior mean}}$$

$$= w\,\hat{\theta}_{MLE} + (1-w)\,\hat{\theta}_{\text{prior mean}}$$

$$w = \frac{n}{n+\alpha+\beta}$$

n: Actual observed shot attempts

$\alpha+\beta$: "Imagined" shot attempts

- Poisson with gamma prior

The posterior mean

$$
\begin{aligned}
E[\lambda \mid y_1, \ldots y_n] &= \frac{a + \sum y_i}{b + n} \\
&= \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n} \\
&= (1-w)\frac{a}{b} + w\hat{\lambda}_{\text{MLE}}
\end{aligned}
$$

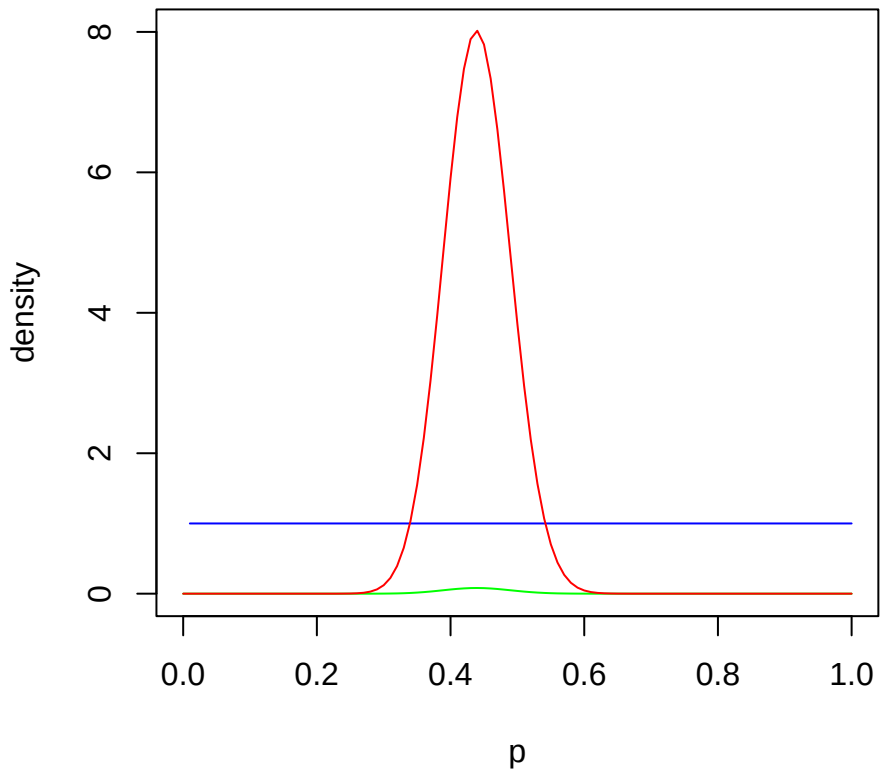
- $w \to 1$ as $n \to \infty$ (data dominates prior)
- b can be interpreted as the number of *prior* observations
  - Analogous to n or total prior exposure
- a can be interpreted as the sum of the counts from prior total exposure of b
  - Analogous to $\sum_i y_i$

Posterior mean and MAP estimates

An early study concerning the sex of newborn Germany babies found that of a total of 98 births, 43 were female. Assume we are using the uniform prior.

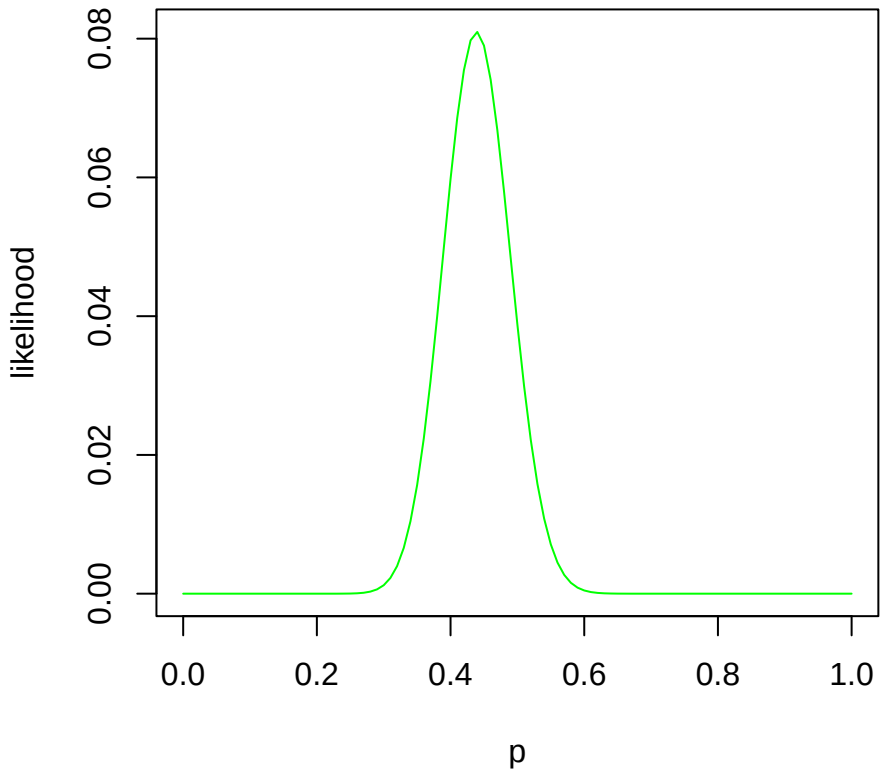
- Plotting the prior distribution (in blue lines), binomial likelihood (in green lines) and the posterior distribution (in red lines).

```
# prior
curve(p/p, from = 0, to = 1, xname = "p", xlab = "p", ylab = "density",
      ylim = c(0, 8), col = 'blue')

# likelihood
curve(choose(98, 43) * p^43 * (1-p)^(98-43),
      from = 0, to = 1,  xname = "p",
      xlab = "p", ylab = "likelihood", add = TRUE,  ylim = c(0, 8), col = 'green')

# posterior
a_post = 1 + 43
b_post = 1 + 98 - 43
curve(gamma(a_post + b_post)/gamma(a_post)/gamma(b_post) *
      p^(a_post - 1) * (1-p)^(b_post - 1), from = 0, to = 1, xname = "p",
      xlab = "p", ylab = "density", add = TRUE, ylim = c(0, 8), col = 'red')
```



It helps to plot the likelihood alone.

```
# likelihood
curve(choose(98, 43) * p^43 * (1-p)^(98-43),
      from = 0, to = 1,  xname = "p",
      xlab = "p", ylab = "likelihood", col = 'green')
```



The posterior probability distribution contains all the current information about the parameter θ. A graphical report on the entire posterior distribution is definitely meaningful and useful. For many practical cases, however, various numerical summaries of the distributions are desirable.

- As for all distributions, the mean of the distribution is an important location summary. What is the mean of the posterior distribution of our above example?

Solution: Following the above formula, we can see the posterior is a $Beta(44, 56)$ distribution. Therefore by the formula of Beta distribution we know the mean is $\frac{44}{(44+56)} = 0.44$.

- Meanwhile, the MAP (maximum a posteriori), which is the mode of the posterior distribution, can be interpreted as the single "most likely" value of the parameter, given the data and the model. What is the MAP estimate in our posterior distribution?

Solution: Consider the density of the posterior distribution, which is

$$\frac{\Gamma(44+56)}{\Gamma(44)\Gamma(56)} p^{44-1}(1-p)^{56-1}.$$

Since the MAP is the mode of the posterior distribution, it is the MLE of p. To maximize the above density, it is equivalent to maximizing the log-likelihood after simplification, which is

$$43\log(p) + 55\log(1-p).$$

By taking the derivative with respect to p and setting it to 0, we get the MAP of p is $\frac{43}{98}$.

Alternatively, you can check the formula for the mode of the Beta distribution directly.

- Is there always a unique MAP value in the posterior distribution?

Solution: No. The posterior distribution might be bimodal or multimodal, therefore we might encounter multiple MAPs.