ST4234: Bayesian Statistics

Tutorial 1 Solution, AY 19/20

Some Comments

- When plotting probability distributions, one should note whether the distribution is discrete or continuous. For discrete distributions, use `type="h"` in the `plot` command. In particular, the points should not be joined together. For continuous distributions, use `type="l"`. When setting up the grid of x-values, the gap between successive points should be sufficiently small so as to obtain a smooth curve for continuous distributions.

- For 1(e), it is sufficient to recognize that the posterior is a beta distribution and to identify the parameters correctly. It is not necessary to work out the normalization constant.

- For 2(a), the aim of this question is to use the fact that for a Beta(a_0, b_0) prior, $a_0 + b_0$ can be interpreted as the "prior sample size" and a_0 the number of ones out of this "prior sample". See Lecture Notes 2 page 25. Usually the "prior sample size" is (much) smaller than the actual sample size because one is not very confident about the value of θ before observing the data or wants to be vague when specifying the prior. In this case the prior sample size is taken to be 25% of the actual sample size.

- For 3(c) , the predictive distribution is not a standard probability distribution so we will have to compute the mean and variance from first definitions.

- For 3(d), it is not advisable to use the distribution with a plug-in estimate for θ to make predictions simply because it has a smaller variance. This is because using a plug-in estimate does not account for uncertainty in estimating θ and will result in predictions that are over confident.

Solutions

1. (a) $Y_i|\theta \sim$ Bernoulli(θ) for $i = 1, \ldots, n$ and $E(Y_i|\theta) = \theta$.
Since $Y_1, \ldots, Y_{100}$ are independent conditional on θ,

$$\begin{aligned} \mathrm{P}(Y_1 = y_1, \ldots, Y_{100} = y_{100}|\theta) &= \prod_{i=1}^{100} \mathrm{P}(Y_i = y_i|\theta) \\ &= \prod_{i=1}^{100} \theta^{y_i}(1-\theta)^{1-y_i} \\ &= \theta^{\sum_{i=1}^{100} y_i}(1-\theta)^{100-\sum_{i=1}^{100} y_i}. \end{aligned}$$

As $Y \sim$ Binomial$(100, \theta)$, $\mathrm{P}(Y = y|\theta) = \binom{100}{y}\theta^y(1-\theta)^{100-y}$.

(b) We can compute and plot the probabilities in `R` using the code below.

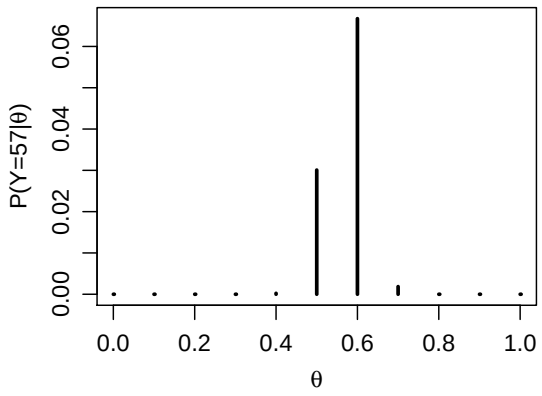
```
theta <- seq(from=0,to=1,by=0.1)
y <- 57
n <- 100
likelihood <- dbinom(y,n,theta)
round(likelihood,4)
plot(theta,likelihood,type="h",lwd=2.5,xlab=expression(theta),
    ylab=expression(paste("P(Y=57|",theta,")")))
```

The table below shows the values of $\mathrm{P}(Y = 57|\theta)$ for each value of θ and the figure below plots these values.

(c) Let $a_1, \ldots, a_{11}$ denote the 11 θ-values. The prior information is $\mathrm{P}(\theta = a_k) = 1/11$ for $k = 1, \ldots 11$. From Bayes' Theorem,

$$\begin{aligned} \mathrm{P}(\theta = a_k|Y = 57) &= \frac{\mathrm{P}(Y = 57, \theta = a_k)}{\mathrm{P}(Y = 57)} \\ &= \frac{\mathrm{P}(Y = 57|\theta = a_k)\mathrm{P}(\theta = a_k)}{\sum_{k'=1}^{11} \mathrm{P}(Y = 57, \theta = a_{k'})} \\ &= \frac{\mathrm{P}(Y = 57|\theta = a_k)\mathrm{P}(\theta = a_k)}{\sum_{k'=1}^{11} \mathrm{P}(Y = 57|\theta = a_{k'})\mathrm{P}(\theta = a_{k'})} \\ &= \frac{\mathrm{P}(Y = 57|\theta = a_k)}{\sum_{k'=1}^{11} \mathrm{P}(Y = 57|\theta = a_{k'})}. \end{aligned}$$

	θ	$\mathrm{P}(Y=57\|\theta)$
1	0.00	0.0000
2	0.10	0.0000
3	0.20	0.0000
4	0.30	0.0000
5	0.40	0.0002
6	0.50	0.0301
7	0.60	0.0667
8	0.70	0.0019
9	0.80	0.0000
10	0.90	0.0000
11	1.00	0.0000

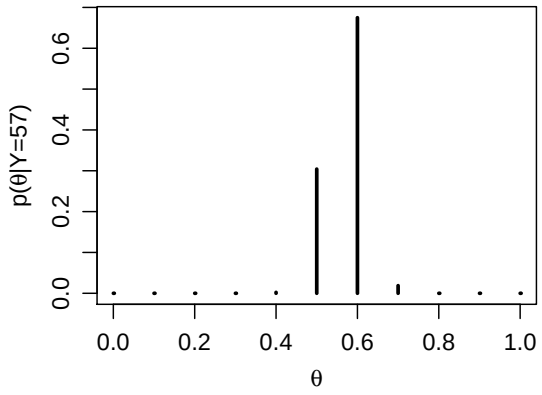


We can compute and plot the posterior probabilities in `R` using the code below.

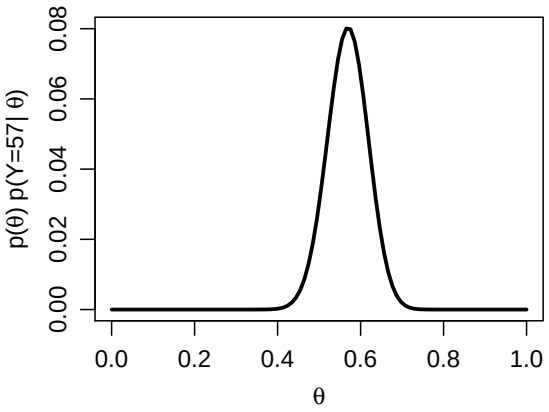
```
posterior <- likelihood/sum(likelihood)
plot(theta,posterior,type="h",lwd=2.5,xlab=expression(theta),
    ylab=expression(paste("p(",theta,"|Y=57)")))
```

The table below shows the values of $\mathrm{P}(\theta|Y=57)$ for each value of θ and the figure below plots this posterior distribution.

	θ	$\mathrm{P}(\theta\|Y=57)$
1	0.00	0.0000
2	0.10	0.0000
3	0.20	0.0000
4	0.30	0.0000
5	0.40	0.0023
6	0.50	0.3041
7	0.60	0.6749
8	0.70	0.0187
9	0.80	0.0000
10	0.90	0.0000
11	1.00	0.0000

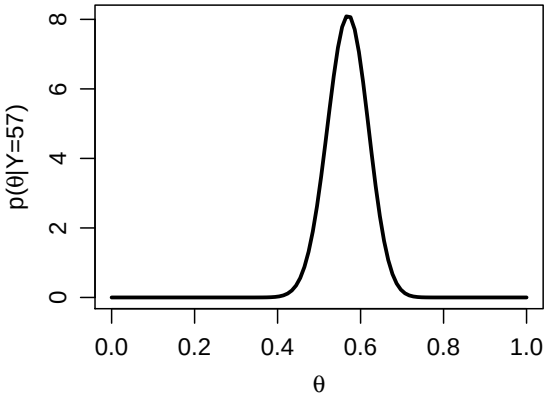


(d) The figure below plots $p(\theta)\mathrm{P}(Y=57|\theta)$ as a function of θ.
Note that here θ can be any value in [0,1]. Hence, we have a continuous curve. The figure can be obtained using the R-code below.




```
theta2 <- seq(from=0, to=1, length.out=100)
plot(theta2,dbinom(y,n,theta2),type="l",lwd=2.5,
    xlab=expression(theta),
    ylab=expression(paste("p(",theta,") p(Y=57|",theta,")")))
```

(e) The posterior distribution of θ is $\text{Beta}(1+57, 1+100-57) = \text{Beta}(58, 44)$ and this distribution is shown in the figure below.



The figure can be obtained using the R-code below.

```
plot(theta2,dbeta(theta2,58,44),type="l",lwd=2.5,
    xlab=expression(theta),
    ylab=expression(paste("p(",theta,"|Y=57)")))
```

In 1(b) and (c) we consider a **discrete** prior for θ that is uniform across the values $\{0.00, 0.10, \ldots, 1.00\}$. The likelihood in 1(b) and the posterior in 1(c) have the same shape but different scale. They have the same shape because the prior is uniform and thus the likelihood is proportional to the posterior as a function of θ. However, they have different scale because the posterior is equal to the likelihood divided by a normalization constant which does not depend on θ. This constant ensures that the posterior sums to 1 over all possible values of $\theta \in \{0.00, 0.10, \ldots, 1.00\}$.

In 1(d) and 1(e), we consider a **continuous** prior for θ that is uniform on [0,1]. The likelihood computed in 1(b) coincides with the values of $p(\theta)\mathrm{P}(Y = 57|\theta)$ computed in 1(d) at the values $\{0.00, 0.10, \ldots, 1.00\}$ since $p(\theta) = 1$. The likelihood in 1(d) and the posterior in 1(e) have the same shape but different scale and the reasoning is the same as before except that in this case, the normalization constant ensures that the posterior integrates to 1 over [0,1] as θ is continuous.

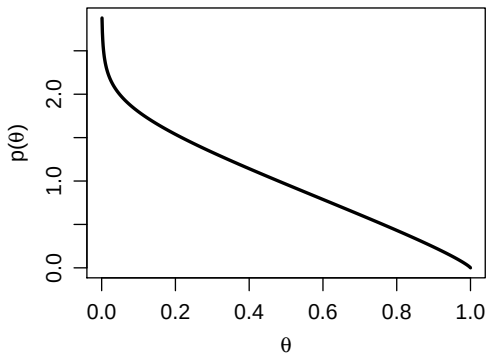
2. Let θ denote the proportion of females in the organization and Y denote the number of females out of the first 11 members. Then $Y \sim \text{Binomial}(11, \theta)$.

   (a) Let us consider the conjugate $\text{Beta}(a_0, b_0)$ prior, which simplifies posterior calculations. The prior information can be represented as

   $$\begin{aligned} \text{prior sample size: } & a_0 + b_0 = 11/4 \\ \text{prior number of 1's: } & a_0 = 1/3 \times 11/4 = 11/12 \\ & \text{This implies that } b_0 = 11/4 - 11/12 = 11/6. \end{aligned}$$

   The R-code below plots the $\text{Beta}(11/12, 11/6)$ prior distribution which is shown in the figure below.

```
a0 <- 11/12
b0 <- 11/6
theta <- seq(from=0,to=1, by=0.001)
plot(theta,dbeta(theta,a0,b0), type="l",lwd=2.5,
    xlab=expression(theta),
    ylab=expression(paste("p(",theta,")")))
```



The standard deviation is the square root of the variance, which is

$$\sqrt{\frac{\frac{11}{12}\cdot\frac{11}{6}}{(\frac{11}{12}+\frac{11}{6})^2(\frac{11}{12}+\frac{11}{6}+1)}} = 0.243.$$

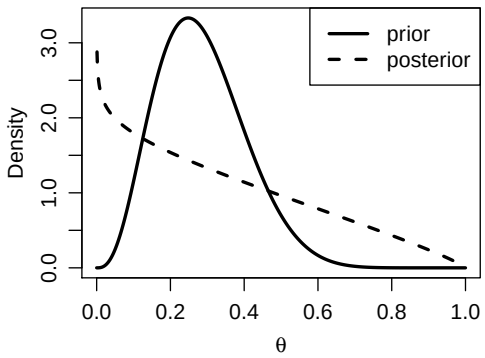
(b) If $y = 3$, then the posterior distribution is Beta $\left(\frac{11}{12}+3, \frac{11}{6}+11-3\right)$ = Beta$(47/12, 59/6)$ = Beta$(3.92, 9.83)$.

```
> y <- 3
> n <- 11
> (a <- a0+y)
[1] 3.916667
> (b <- b0+n-y)
[1] 9.833333
```

The R-code below plots this posterior distribution which is shown in the figure below. The prior is also shown in dotted lines.

```
> plot(theta, dbeta(theta,a,b),type="l",lwd=2.5,
+     xlab=expression(theta), ylab="Density")
> points(theta, dbeta(theta,a0,b0),type="l",lwd=2.5,lty=2)
> legend("topright",legend=c("prior","posterior"),lty=c(1,2),
+     lwd=2.5,cex=1)
```

The mean is 0.285, the median is 0.274 and the mode is 0.248 (computed in R).




```
> (mean <- a/(a+b))
[1] 0.2848485
> (median <- qbeta(0.5, a, b))
[1] 0.2741738
> (mode <- (a-1)/(a+b-2))
[1] 0.248227
```

Alternatively, you can use the formulas to compute the mean and the mode from the posterior Beta$(47/12, 59/6)$:

$$\text{mean} = \frac{a}{a+b} = \frac{\frac{47}{12}}{\frac{47}{12}+\frac{59}{6}} = \frac{47}{165} = 0.2848485,$$
$$\text{mode} = \frac{a-1}{a+b-2} = \frac{\frac{47}{12}-1}{\frac{47}{12}+\frac{59}{6}-2} = \frac{35}{141} = 0.248227.$$

The median cannot be computed analytically.

(c) The 50% highest posterior density region is (0.174, 0.334). The 50% quantile-based confidence interval is (0.198,0.361). We can compute these intervals using the R-code below.

```
> require(TeachingDemos)
> (hpdr <- hpd(qbeta,shape1=a,shape2=b,conf=0.5))
[1] 0.1738371 0.3342303
> hpdr[2] - hpdr[1]
[1] 0.1603932
```

```
>
> # quantile-based CI
> (CI <- qbeta(c(0.25,0.75),a,b))
[1] 0.1975017 0.3611453
> CI[2] -CI[1]
[1] 0.1636436
```

The quantile-based confidence interval is slightly wider than the highest posterior density region and is shifted to the right of the highest posterior density region.

(d) It is not surprising as this value ($86/433 \approx 0.199$) is not lying in the tails of the posterior density. It falls inside the 50% highest posterior density region.

3. (a) Let Y_1 denote the number of children out of 15 who tested positive for the disability. Then $Y_1 \sim \text{Binomial}(15, \theta)$. Using a uniform prior distribution, the posterior distribution of θ is $\text{Beta}(1 + y_1, 1 + n_1 - y_1) = \text{Beta}(3, 14)$.

```
> a0 <- 1; b0 <- 1
> y1 <- 2; n1 <- 15
> (a <- a0+y1)
[1] 3
> (b <- b0+n1-y1)
[1] 14
```

(b) i. We need to assume that Y_2 is **conditionally independent** of Y_1 given θ.

ii. Note that $Y_2|\theta \sim \text{Binomial}(n_2, \theta)$.

$$\begin{aligned}
\text{P}(Y_2 = y_2|Y_1 = 2) &= \int_0^1 P(Y_2 = y_2|\theta)p(\theta|Y_1 = 2)\,\text{d}\theta \\
&= \int_0^1 \binom{n_2}{y_2}\theta^{y_2}(1-\theta)^{n_2-y_2}\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}\,\text{d}\theta \\
&= \frac{\binom{n_2}{y_2}}{B(a,b)}\int_0^1 \theta^{y_2+a-1}(1-\theta)^{n_2-y_2+b-1}\,\text{d}\theta,
\end{aligned}$$

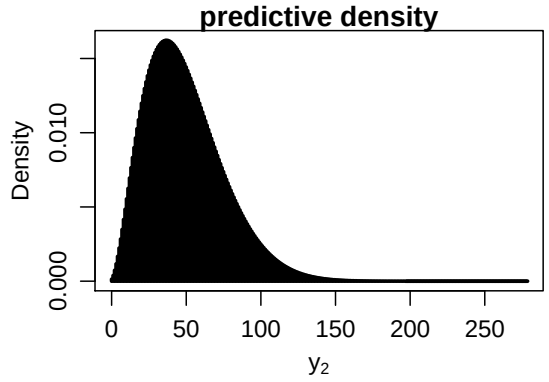
where $a = 3$, $b = 14$.

iii. Therefore

$$\mathrm{P}(Y_2 = y_2 | Y_1 = 2) = \frac{\binom{n_2}{y_2} B(y_2 + a, n_2 - y_2 + b)}{B(a,b)} \underbrace{\int_0^1 \frac{\theta^{y_2+a-1}(1-\theta)^{n_2-y_2+b-1}}{B(y_2+a, n_2-y_2+b)} \, \mathrm{d}\theta}_{=1}$$

$$= \frac{\binom{n_2}{y_2} B(y_2 + a, n_2 - y_2 + b)}{B(a,b)}.$$

(c) The figure below plots $\mathrm{P}(Y_2 = y_2 | Y_1 = 2)$ as a function of y_2.

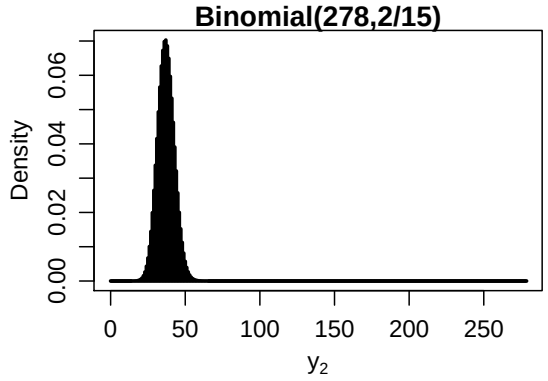


Note that $y_2 \in \{0, 1, \ldots, 278\}$. The mean of Y_2 given $Y_1 = 2$ is 49.06 and the standard deviation is 25.73.

```
> n2 <- 278
> y2 <- seq(from=0, to=n2, by=1)
> predictive_density <- function(y2, n2, a, b){
+   choose(n2,y2)*beta(y2+a,n2-y2+b)/beta(a,b)
+  }
> y2prob <- predictive_density(y2, n2, a, b)
> plot(y2,y2prob,type="h",lwd=2.5,ylab="Density",
+     xlab=expression(y[2]),main="predictive density")
> # mean
> (post.mean <- sum(y2*y2prob))
[1] 49.05882
> # standard deviation
> post.var <- sum(y2^2*y2prob) - post.mean^2
```

```
> (post.sd <- sqrt(post.var))
[1] 25.73196
```

(d) The distribution of $\mathrm{P}(Y_2 = y_2|\theta = \hat{\theta})$ is Binomial$(278, 2/15)$ and it is plotted in the figure below.



The mean of Y_2 given $\theta = \hat{\theta}$ is 37.07 and standard deviation is 5.67.

```
> plot(y2,dbinom(y2,n2,2/15),type="h",lwd=2.5,ylab="Density",xlab=expression(y[2]))
> 278*2/15
[1] 37.06667
> sqrt(278*2/15*13/15)
[1] 5.667843
```

The distribution $\mathrm{P}(Y_2 = y_2|\theta = \hat{\theta})$ is more peaked and the mean and standard deviation of Y_2 given $\theta = \hat{\theta}$ are both lower than that of the posterior predictive distribution. The posterior predictive density is preferable since the plug-in density ignores uncertainty about θ. As the sample size in the pilot study ($n_1 = 15$) is small, it is important to take into account the high variability in θ.

Remark: $\mathrm{P}(Y_2 = y_2|\theta = \hat{\theta})$ is not $\mathrm{P}(Y_2 = y_2|\hat{\theta})$. $\mathrm{P}(Y_2 = y_2|\theta = \hat{\theta})$ is the conditional probability of $Y_2 = y_2$ given $\theta = \hat{\theta}$. It is calculated from the conditional distribution $Y_2|\theta$, where θ is a random variable (that is conditioned on) and $\hat{\theta}$ is the value of θ. In contrast, $\mathrm{P}(Y_2 = y_2|\hat{\theta})$ comes from the conditional distribution $Y_2|\hat{\theta}$, where $\hat{\theta}$ is treated as a random variable. Since $\hat{\theta} = Y_1/n_1$, $Y_2|\hat{\theta}$ is the same as $Y_2|Y_1$. So $\mathrm{P}(Y_2 = y_2|\theta = \hat{\theta}) = \mathrm{P}(Y_2 = y_2|Y_1 = 2)$, as calculated in Part (b)(iii).