STAT314 - 2017S2: Solutions to Ass1

Elena Moltchanova

August 2017

Problem 1:

Let A and B denote the events of a panda belonging to species A and B respectively. $Pr(A) = Pr(B) = 0.5$. Let T_k and T_k^C denote the events a panda having twins or a singleton respectively as the result of birth number k. $Pr(T_k|A) = 1 - Pr(T_k^C|A) = 0.1$ and $Pr(T_k|B) = 1 - Pr(T_k^C|B) = 0.3$.

(a)

$$\begin{aligned} Pr(A|T_1) &= \frac{Pr(T_1|A)Pr(A)}{Pr(T_1|A)Pr(A) + Pr(T_1|B)Pr(B)} \\ &= \frac{.1 * .5}{.1 * .5 + .3 * .5} = 0.25 \end{aligned}$$

(b)

$$\begin{aligned} Pr(T_2|T_1) &= Pr(T_2, A|T_1) + (T_2, B|T_1) \\ &= Pr(T_2|A, T_1)Pr(A|T_1) + Pr(T_2|B, T_1)Pr(B|T_1) \\ &= Pr(T_2|A)Pr(A|T_1) + Pr(T_2|B)(1 - Pr(A|T_1)) \\ &= 0.1 * 0.25 + 0.3 * 0.75 = 0.25 \end{aligned}$$

(c)

$$\begin{aligned} Pr(A|T_1, T_2^C) &= \frac{Pr(T_2^C|A)Pr(A|T_1)}{Pr(T_2^C|A)Pr(A|T_1) + Pr(T_2^C|B)Pr(B|T_1)} \\ &= \frac{0.9 * 0.25}{0.9 * 0.25 + 0.7 * 0.75} = 0.3 \end{aligned}$$

Problem 2:

Given a coin with probability p of obtaining *heads* in any random toss, the probability of obtaining y *heads* out of two tosses is $\binom{2}{y} p^y (1-p)^{2-y}$. For the fair coin $p = 0.5$, for the other one $p = 1$.

(a)

$$\begin{aligned} Pr(Fair|y = 2) &= \frac{Pr(y = 2|Fair)Pr(Fair)}{Pr(y = 2|Fair)Pr(Fair) + Pr(y = 2|NotFair)Pr(NotFair)} \\ &= \frac{0.5^2 * 0.5}{0.5^2 * 0.5 + 1^2 * 0.5} = 0.2 \end{aligned}$$

(b) There are several ways to look at the problem. The first one is to simply test $H_0 : p = 0.5$. The definition of the p-value is the probability of obtaining an observation which is as least as extreme as what we have (i.e., more extreme than 2 *heads*) if the null hypothesis is actually true. For a two-sided test, $p = 2 * 0.5^2 = 0.5$.

For a one-sided test, which makes more sense because we know that the alternative coin is one-sided, $p = 0.25$.

The important thing to notice is that while in classical statistics, p-value is $Pr(data|H_0)$, in Bayesian statistics we can obtain $Pr(H_0|data)$, which is what we are actually after.

Problem 3:

For n i.i.d. observations from geometric distribution, the joint likelihood will be

$$p(x_1, ..., x_n|p) = (1-p)^{n\bar{x}-n}p^n$$

(a) The log-likelihood will thus be

$$L = n(\bar{x}-1)\log(1-p) + n\log(p)$$

Differentiating with respect to p and setting the derivative to 0 we obtain

$$\frac{dL}{dp} = -\frac{n(\bar{x}-1)}{1-p} + \frac{n}{p} = 0$$

$\implies \hat{p} = \frac{1}{\bar{x}}$.

To check that it is, indeed, a maximum, obtain the second derivative:

$$\frac{d^2L}{dp^2} = -\frac{n(\bar{x}-1)}{(1-p)^2} - \frac{n}{p^2}.$$

Note, that because $\bar{x} > 1$ and $0 < p < 1$, the second derivative is always negative, and thus $\hat{p}$ is the MLE of p. (You can also evaluate the second derivative specifically for $\hat{p}$ and check that it is negative.)

(b) Let $p \sim Beta(a, b)$, i.e., $p(p) = \frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}$. Then, using Bayes formula, we obtain

$$\begin{aligned} p(p|x) \propto p(x|p)p(p) &\propto p^{n\bar{x}-n}(1-p)^n p^{a-1}(1-p)^{b-1} \\ &= p^{a+n-1}(1-p)^{b+n\bar{x}-n-1}. \end{aligned}$$

Which implies that $p|x \sim Beta(a+n, b+n\bar{x}-n)$.

(c) The posterior mean of p is $E(p|x) = \frac{a+n}{a+b+n\bar{x}}$. When $n \to \infty$, $E(p|x) \to \frac{1}{\bar{x}}$, i.e., Bayesian posterior estimate approaches the frequentist MLE.

Problem 4.

(a) Let the numebr of days a week a person exercises have a binomial distribution $y_i \sim Bin(7, p)$ and assign a uniform prior $p \sim Beta(1, 1)$. We know, that in this case the posterior distribution will be $p|y \sim Beta(1 + \sum_i y_i, 1 + 7 * n - \sum_i y_i)$, i.e., $p|y \sim Beta(1 + 141, 1 + 700 - 141)$.

```
post.sample <- rbeta(10^3,142,560)
# posterior mean:
mean(post.sample)
```

```
## [1] 0.2022028
```

```
# posterior 95% credible interval
quantile(post.sample,c(0.025,.975))
```

```
##      2.5%     97.5%
## 0.1733330 0.2325773
```

Based on the observations, we estimate the probability of a person exercising on any particular day to be on average 0.2 and to lie with 95% probability within the interval (0.17,0.23)

(b)

```
post.pred.sample <- rbinom(10^3,7,post.sample)
mean(post.pred.sample>=3)
```

```
## [1] 0.149
```

The posterior predictive probability of exercising at least three days a week is 0.15.

(c)

```
post.sample2 <- rbeta(10^3,302+141,458+700-141)
mean(post.sample2)
```

```
## [1] 0.30317
```

```
quantile(post.sample2,c(0.025,.975))
```

```
##      2.5%     97.5%
## 0.2799550 0.3285851
```

```
post.pred.sample2 <- rbinom(10^3,7,post.sample2)
mean(post.pred.sample2>=3)
```

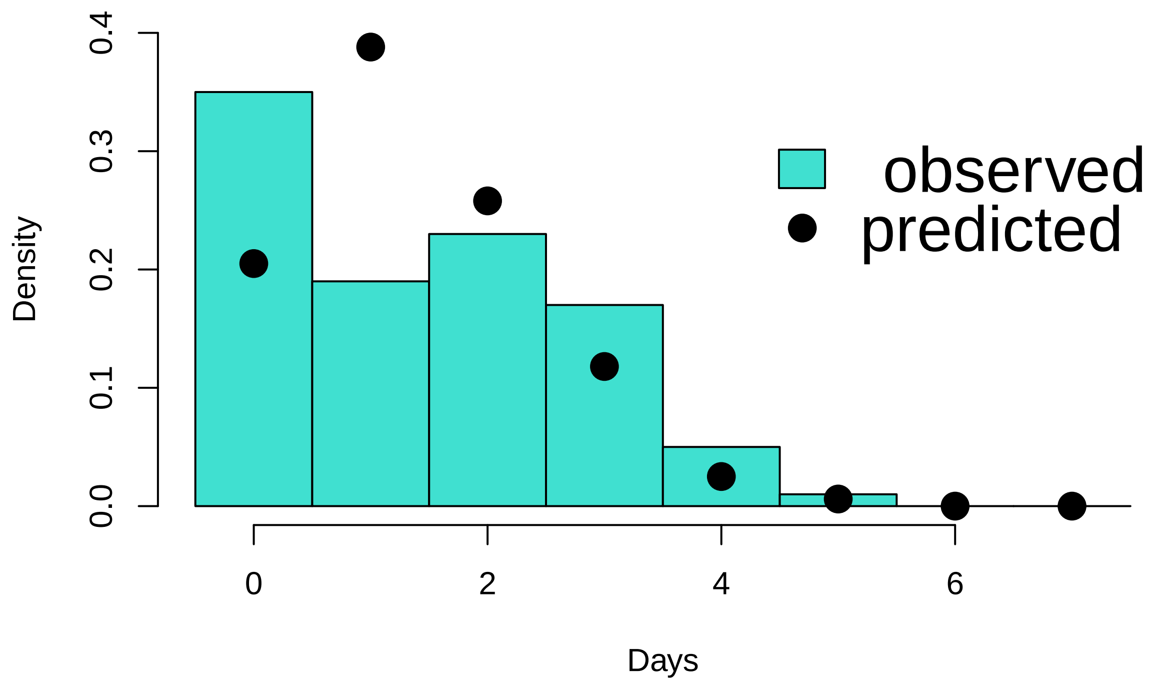
```
## [1] 0.348
```

(d)

```
y <- rep(0:5,c(35,19,23,17,5,1))
hist(y,seq(-.5,7.5,1),col='turquoise',freq=F,main='',xlab='Days',ylim=c(0,.4))

points(0:7,table(factor(post.pred.sample,0:7))/10^3,pch=16,cex=2)

legend(4,.35,fill='turquoise','observed',bty='n',cex=2)
legend(4.2,.3,pch=16,cex=2,'predicted',bty='n')
```



Does posterior predictive distribution looks like the observed one? If it does not, than the model does not fit the observations.