HOMEWORK 6, STAT 251

Do the following using R. You must also turn in a copy of your R code.

Table 1. Career Major League Baseball Statistics for Ted Williams and Joe DiMaggio

Player	At Bats	Hits	Batting Average	Home runs	Home run average
Ted Williams	7,706	2,654	.3444	521	.0676
Joe DiMaggio	6,821	2,214	.3246	361	.0529

(1) During the 1940s and 1950s, Ted Williams played baseball for the Boston Red Sox and Joe DiMaggio played for the New York Yankees. They were regarded as among the best players of their day, and the question of who was better was the subject of many heated arguments among baseball aficionados. Williams had the better batting record. Was he in fact a better hitter? Their major league career statistics are given in Table 1. Assume that at-bats are exchangeable for each player. (This assumption is only approximately correct for me. Players learn, change their batting style from time to time, and get older. Moreover, they do not always play in the same ball park, against the same pitcher, in the same weather, the style of ball may change, and so on. But the assumption may not be far off. And if we do not assume something, we will not be able to draw any conclusions!)

(a) First consider hits. A baseball afficionado's prior probabilities concerning these two players' "true proportions" of hits are given by independent beta(a, b) densities, where a = 50 and b = 120. (The values of a and b we assume do not matter much. Both Williams and DiMaggio had so many at-bats that we would get about the same answer even if we were to assume a = b = 1.) Find this fan's posterior probability (given the preceding career statistics) that Ted Williams was a better hitter than Joe DiMaggio. (Interpret better to mean that there was a higher proportion of hits in the Williams model.)]

(b) Find this fan's 95% posterior probability interval for the increase in the hitting success proportion for Williams over DiMaggio (that is, for $\theta_W - \theta_D$).

(c) Plot the (estimated) posterior density of $\theta_W - \theta_D$ (in black) and the (estimated) prior density of $\theta_W - \theta_D$ (in gray) on the same plot. Make sure to properly label all plot components.

(d) Now consider home runs. Find this fan's posterior probability that Williams was a better home-run hitter than DiMaggio, assuming that for each player his prior distribution for the chance that player would hit a home run in any given at-bat is given by independent beta(20, 500) densities.

(e) Based on the answers to the previous four parts, what conclusion(s) if any should the fan draw in comparing the two baseball players? Be sure to thoroughly explain your answer.

(2) To reinforce your understanding of prior-predictive distributions and posterior-predictive distributions we now consider predictions from the same baseball example. In baseball, one of the revered accomplishments is to have a batting average of .400 or above (baseball statistics always use three decimal places for a batting average) over the course of a season. This means that a player gets a hit in at least 40.0% of their at-bats during that season. As before, let θ_W represent the "true proportion" of hits for Ted Williams. We assume that Williams' "true" but unobservable probability of getting a hit, θ_W, was constant over the course of his career (though, as discussed previously, there are reasons to doubt this assumption). **Throughout all parts of this problem of the homework, we assume a $Beta(50, 120)$ prior distribution for θ_W.**

(a) Suppose we knew that Williams would have five at bats in his *first* career game. What is the prior-predictive probability that Williams would have at least a .400 (recorded) batting average after this first game? (That is, what is the prior predictive probability he gets at least two hits in his first five at bats, so that his recorded average is at least 2/5 =.400?) Recall that we assume the at-bats are conditionally iid, so that $Y_{new,\text{ first game}}|\theta_W \sim Binomial(n_{new,\text{ first game}} = 5, \theta_W)$. Calculate the probability EXACTLY.

(b) Refer to the previous question. Now, ESTIMATE the requested probability by appropriately using Monte Carlo.

(c) Suppose we knew that Williams would have 605 at bats in his *first* career season. What is the prior-predictive probability that Williams would have at least a .400 (recorded) batting average after this first season? Calculate this probability exactly.

(d) Refer to the previous question. Now, ESTIMATE the requested probability by appropriately using Monte Carlo.

(e) The table of Williams' actual batting career is provided above. We will suppose he had the chance to play one extra game right after this career, and that he would have five at-bats. Given our prior beliefs about θ_W and then the observed data from the table below, what is the probability that Williams would get at least two hits if he were to play one extra game at the end of his career, and we knew that he would have five at-bats? Calculate this posterior probability exactly.

(f) Refer to the previous question. Now, ESTIMATE the requested probability by appropriately using Monte Carlo.

(g) We will suppose he had the chance to play one extra *season* right after his career, and that he would have 605 at-bats. Given our prior beliefs about θ_W and then the observed data from the table below, what is the probability that Williams would have a recorded batting average of at least .400 in this new season? Calculate this probability exactly. (Hint: we need the probability he would get at least 242 hits in the 605 new at-bats.)

(h) Refer to the previous question. Now, ESTIMATE the requested probability by appropriately using Monte Carlo.

(i) Repeat the previous **eight** questions for Joe DiMaggio. We will use the same prior distribution for θ_D (i.e., the Beta(50, 120) distribution), but the posterior will be different because of DiMaggio's different batting record.

(j) What is the probability that at the beginning of their careers, (so using *prior*-predictive distributions) that Williams' recorded batting average would be strictly better than DiMaggio's if each had 605 at bats in their first season? Use Monte Carlo.

(k) What is the probability that after their careers, (so using *posterior*-predictive distributions) that Williams' recorded season average in a new season with 605 at bats would be strictly better than DiMaggio's recorded season average in a new season with 605 at bats?