Math 390.03-02 / 650.03-01 Spring 2016
Midterm Examination One Solutions

Professor Adam Kapelner

Monday, March 14, 2016

Full Name _____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using a cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

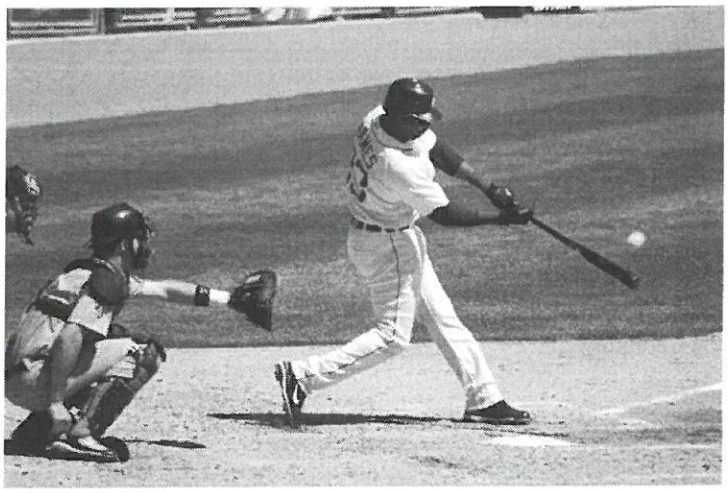
_____ _____

signature date

Instructions

This exam is seventy five minutes and closed-book. You are allowed one page (front and back) of a "cheat sheet." You may use a graphing calculator of your choice. Please read the questions carefully. If the question reads "compute," this means the solution will be a number otherwise you can leave the answer in *any* widely accepted mathematical notation which could be resolved to an exact or approximate number with the use of a computer. I advise you to skip problems marked "[Extra Credit]" until you have finished the other questions on the exam, then loop back and plug in all the holes. I also advise you to use pencil. The exam is 100 points total plus extra credit. Partial credit will be granted for incomplete answers on most of the questions. $\boxed{\text{Box}}$ in your final answers. Good luck!

Problem 1 This question is about "batting averages" in baseball.



Every hitter's *sample* batting average (BA) is defined as:

$$BA := \frac{\text{sample \# of hits}}{\text{sample \# of at bats}}$$

In this problem we care about estimating a hitter's *true* batting average which we call θ. Each player has a different θ. In order to estimate the true batting average, we use the sample batting average as defined above.

(a) [2 pt / 2 pts] For the remainder of the problem, we assume that each at bat for any player are *conditionally* $\overset{iid}{\sim}$ based on the players' true batting average, θ. So if a player has n at bats, then each successful hit in each at bat can be modeled via

$$X_1 \mid \theta,\ X_2 \mid \theta, \ldots,\ X_n \mid \theta \overset{iid}{\sim} \text{Bernoulli}\,(\theta)\,.$$

Under this model above, if the player had $n = 4$ at bats, would the $\mathbb{P}(X_3, X_2, X_4, X_1)$ be equal to the $\mathbb{P}(X_1, X_2, X_3, X_4)$? Yes / no.

Yes

(b) [3 pt / 5 pts] If the player had $n = 4$ at bats and $\sum_{i=0}^{n} x_i = 0$ hits, compute $\hat{\theta}_{\text{MLE}}$.

$\hat{\theta}_{MLE} = \bar{x} = \frac{0}{4} = 0$

(c) [1 pt / 9 pts] Compute a frequentist confidence interval for θ given the data in (b).

$CI_{\theta,95\%} = \{0\}$ since $SE(\hat{\theta}_{MLE}) = 0$

(d) [3 pt / 12 pts] Describe in English the main problem with the interval in (c).

It is not an interval and thereby cannot satisfy the def. of a frequentist CI.

(e) [2 pt / 14 pts] Set the following prior: $\theta \sim \mathrm{U}(0, 1)$. Is this an informative prior for the true batting average? Yes/no

No. $U(0,1)$ is an "objective" or "uninformative" prior.

(f) [5 pt / 19 pts] Given the prior in (e) and the data in (b), find the posterior distribution of this player's true batting average.

$\theta \sim Beta(1,1) = U(0,1)$

$\theta|x \sim Beta(\alpha+x, \beta+n-x) = Beta(1+0, 1+4-0) = Beta(1,5)$

(g) [3 pt / 22 pts] Based on your posterior distribution in (f), give your best estimate to the value of θ which minimizes squared error loss.

$\hat{\theta}_{MMSE} = E(\theta|x) = \frac{\alpha+x}{\alpha+\beta+n} = \frac{1+0}{1+1+4} = \boxed{\frac{1}{6}}$

(h) [2 pt / 24 pts] Based on your posterior distribution in (f), describe ~~in some way~~ with an integral or the R language your best estimate to the value of θ which minimizes absolute error loss but do not compute.

$\hat{\theta}_{MAE} = Med(\theta|x)$ OR $qbeta(0.5, 1, 5)$ or $\{x : \int_0^x \frac{1}{B(1,5)} \theta^0 (1-\theta)^4 d\theta = 0.5\}$

grids by shade

720 - 784

-5620

(i) [1 pt / 28 pts] Based on your posterior distribution in (f), give your best estimate to the value of θ using the posterior mode.

(f) $\theta|x \sim Beta(1,5)$ $\quad Mode[\theta|x] = \frac{\alpha+x-1}{n+\alpha+\beta-2} = \frac{1+0-1}{4+1+1-2} = \boxed{0}$

(j) [5 pt / 33 pts] Find an integral expression for the probability this hitter bats above a 300 batting average (which means the true batting average is 0.3 or greater). Do not compute.

$$P(\theta > 0.3|x) = \int_{0.3}^{1} \frac{1}{B(1,5)} \theta^0 (1-\theta)^4 \, d\theta$$

(k) [5 pt / 38 pts] Assuming you have access to R and its function `qbeta`, give the 95% credible region for θ. The three arguments for `qbeta` are (1) quantile (2) alpha and (3) beta. Then, provide an interpretation for this interval.

$$CR_{\theta,95\%} = \left[qbeta(0.025, 1, 5), qbeta(0.975, 1, 5) \right]$$

(l) [3 pt / 41 pts] What would the Jeffrey's prior be in ~~this~~ the model situation described in (g)?

$\theta \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right)$

(m) [2 pt / 43 pts] Would the posterior under the Haldane prior be proper given the data in (b)? Yes / No.

$\theta \sim Beta(0,0) \Rightarrow \theta|x \sim Beta(0+0, 0+4-0) = Beta(0,4) \Rightarrow$ No.

(n) [5 pt / 48 pts] The batting average is only measured as the batting average and never logged or transformed. Would there be any value in using the Jeffrey's prior instead of the prior in (e)? Discuss.

No. Since we do not need a prior which is robust to changes in measurement of our parameter of interest. However, the Jeffrey's prior can be useful as an objective prior just like the uniform or Haldane.

(o) [5 pt / 53 pts] Looking at the entire dataset for 6,061 batters who had 100 or more at bats, I fit a beta function to the sample batting averages and estimated $\alpha = 42.3$ and $\beta = 127.7$ (which we called "empirical Bayes" estimates in class). Consider building a prior from this estimate as

$$\theta \sim \text{Beta}\,(42.3,\ 127.7)\,.$$

Would a prior based on these hyperparameter estimates be "objective"? Yes / No. Why?

No. This is a strong prior and hence very informative i.e. non-objective.

(p) [2 pt / 55 pts] Is the prior from (o) considered a "conjugate prior"? Yes / No.

Yes. Any Beta is conjugate.

(q) [3 pt / 58 pts] Using the prior from (o), find the $\hat{\theta}_{\text{MMSE}}$ considering ~~the prior alone~~ when any data whatsoever. Round to 3 digits.

$$\hat{\theta}_{\text{MMSE}} = E(\theta) = \frac{\alpha}{\alpha+\beta} = \frac{42.3}{42.3+127.7} = .249$$

(r) [4 pt / 62 pts] Using the prior from (o) and the data from (b), find the posterior $\hat{\theta}_{\text{MMSE}}$. Round to 3 digits.

$$\hat{\theta}_{\text{MMSE}} = E(\theta|x) = \frac{\alpha+x}{n+\alpha+\beta} = \frac{42.3+9}{4+42.3+127.7} = .243$$

(s) [5 pt / 67 pts] The posterior estimate from (q) is different from the frequentist estimate in (b) due to shrinkage. What is the proportion of shrinkage for the posterior estimate in (q)? We denoted this as ρ in class. Round to 3 digits.

$$\rho = \frac{\alpha+\beta}{\alpha+\beta+n} = \frac{42.3+127.7}{42.3+127.7+4} = .977$$

(t) [1 pt / 71 pts] [Extra Credit] Using the Bayesian CLT, compute a 95% credible region for θ for the data in (b) to the nearest ~~two decimal points~~ three digits. and the prior in (o)

$\theta|x \overset{d}{\approx} N(E[\theta|x], SE[\theta|x]^2) = N(.243$

$SE[\theta|x] = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = \sqrt{\frac{(42.3)(131.7)}{(42.3+131.7)^2(42.3+131.7+1)}} = .0324$

$CR_{\theta,15\%} = [.243 \pm 2 \cdot .0324] = [.178, .308]$

(u) [3 pt / 71 pts] Based on the data in (b) and the prior in (o), what is the probability this batter gets a hit on his next at bat?

$X^* | X \sim Bern\left(\frac{\alpha+x}{\alpha+\beta+n}\right) = Bern\left(\frac{42.3}{42.3+131.7+0}\right) = Bern(.243)$

$\Rightarrow P(X^*|X=1) = .243$

(v) [5 pt / 79 pts] Based on the data in (b) and the prior in (o), write an exact expression for the batter getting 14 or more hits on the next 20 at bats. You can leave your answer in terms of the beta function. Do not compute explicitly.

$X^* | X \sim BetaBin(m, \alpha+x, \beta+n-x) = BernBin(20, 42.3, 174)$

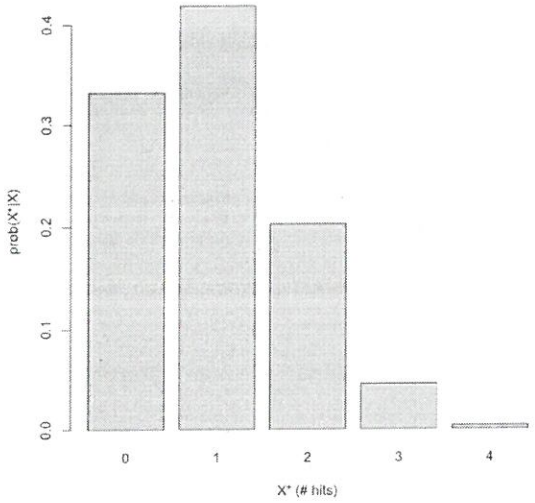
$P(X^* | X \geq 14) = \sum_{x^*=14}^{20} \binom{20}{x^*} \frac{B(x^*+42.3,\ 20-x^*+131.7)}{B(42.3, 131.7)}$

(w) [6 pt / 85 pts] Based on the data in (b) and the prior in (o), find the kernel of the distribution for the number of hits this batter gets in the next m at bats. Partial credit is given.

$P(X^*|X) = BernBin(m, 42.3, 131.7) = \binom{m}{x^*} \frac{B(x^*+42.3,\ m-x^*+131.7)}{B(42.3, 131.7)}$

$\propto \frac{1}{x^*!(m-x^*)!} \frac{\Gamma(x^*+42.3+m-x^*+131.7)}{\Gamma(x^*+42.3)\Gamma(m-x^*+131.7)} \propto x^*!(m-x^*)!\,\Gamma(x^*+0.3)\,\Gamma(m-x^*+0.7)$

(x) [2 pt / 87 pts] Based on the data in (b) and the prior in (o), the joint posterior predictive distribution for $n = 4$ looks like as follows.



Does the data from (b) look abnormal for this model? Yes / no.

(b) $x=0$ Since $P(X^*|x=0) \neq 0 \Rightarrow$ data is good $\Rightarrow$ No

(y) [7 pt / 94 pts] Test the following hypotheses by finding an expression for the Bayesian p-val for the data in (b) and prior in (o):

integral or R/python

$$H_0 : \theta \geq \theta_0$$
$$H_a : \theta < \theta_0$$

where $\theta_0 = \mathbb{E}[\theta]$. That is, we're testing if this batter is truly "below-average" as compared to the 6,061 career major league baseball players from the official dataset.

$\theta_0 = E(\theta) = .249$ (from 2)

$$p_{val} := P(H_0|x) = P(\theta \geq .249|x) = \int_{.249}^{1} \frac{1}{B(82.3, 131.7)} \theta^{81.3}(1-\theta)^{130.7} d\theta$$

$$\text{or} = 1 - pbeta(.249, 82.3, 131.7)$$

or R language

(z) [7 pt / 101 pts] Write an integral expression for K, the Bayes Factor in favor of H_a.

$$K_{H_a,H_0} = \frac{P(X|H_a)}{P(X|H_0)} = \frac{P(X|\theta<.249)}{P(X|\theta\geq.249)} = \frac{\int_0^{.249} \theta^0(1-\theta)^4 \frac{1}{B(42.3,122.7)} \theta^{41.3}(1-\theta)^{126.7}\, d\theta}{\int_{.249}^1 \frac{1}{B(42.3,122.7)} \theta^{41.3}(1-\theta)^{126.7}\, d\theta}$$

$$= \frac{P(\theta<.249|x)/P(\theta<.249)}{P(\theta\geq.249|x)/P(\theta\geq.249)} = \frac{pbeta(.249,42.3,134.7)/pbeta(.249,42.3,122.7)}{1-pbeta(.249,42.3,134.7)/1-pbeta(.249,42.3,122.7)}$$

(aa) [3 pt / 104 pts] For the model in (a), specify $\mathcal{F}$ for the likelihood of a hit at a single at bat.

$\mathcal{F} = \{\theta^x(1-\theta)^{1-x} : \theta\in(0,1)\}$ i.e. a Bernoulli likelihood model

(bb) [3 pt / 107 pts] [Extra credit] For the prior in (o), compute $\mathbb{E}_X[\mathbb{E}_\theta[\theta \mid X]]$.

$E_x(E_\theta(\theta|x)) = E(\theta) = .249$

(cc) [3 pt / 110 pts] [Extra credit] For the data in (b), what is the frequentist predictive distribution?

Beta(a)