Math 390.03-02 / 650.03-01 Spring 2016
Midterm Examination Two

Solutions

Professor Adam Kapelner

Wednesday, April 20, 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

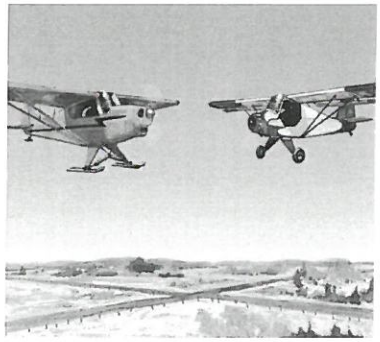
This exam is seventy five minutes and closed-book. You are allowed one page (front and back) of a "cheat sheet." You may use a graphing calculator of your choice. Please read the questions carefully. If the question reads "compute," this means the solution will be a number otherwise you can leave the 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!

This exam assumes you will be able to compute some answers numerically given a computer. Thus, you can leave your answer in terms of any of the following functions as long as the question does not say "compute explicitly." However, you must make clear the numerical values that are parameters for these functions.

Distribution	Quantile Function	CDF function	Sampling Function
beta	qbeta(p, α, β)	pbeta(x, α, β)	rbeta(α, β)
betabinomial	qbetabinom(p, n, α, β)	pbetabinom(x, n, α, β)	rbetabinom(n, α, β)
betanegativebinomial	qbeta_nbinom(p, r, α, β)	pbeta_nbinom(x, r, α, β)	rbeta_nbinom(r, α, β)
binomial	qbinom(p, n, θ)	pbinom(x, n, θ)	rbinom(n, θ)
exponential	qexp(p, θ)	pexp(x, θ)	rexp(θ)
gamma	qgamma(p, α, β)	pgamma(x, α, β)	rgamma($\cancel{n}$, α, β)
geometric	qgeom(p, θ)	pgeom(x, θ)	rgeom(θ)
inversegamma	qinvgamma(p, α, β)	pinvgamma(x, α, β)	rinvgamma($\cancel{n}$, α, β)
negative-binomial	qnbinom(p, r, θ)	pnbinom(x, r, θ)	rnbinom(r, θ)
normal (univariate)	qnorm(p, θ, σ)	pnorm(x, θ, σ)	rnorm(θ, σ)
normal (multivariate)			rmvnorm($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$)
poisson	qpois(p, θ)	ppois(x, θ)	rpois(θ)
T (standard)	qt(p, ν)	pt(x, ν)	rt(ν)
T (nonstandard)	qt.scaled(p, ν, μ, σ)	pt.scaled(x, ν, μ, σ)	rt.scaled(ν, μ, σ)
uniform	qunif(p, a, b)	punif(x, a, b)	runif(a, b)

Table 1: Functions from R (in alphabetical order) that can be used on this exam. The quantile function solves for x in the following equation: $\int_{-\infty}^{x} f(x)dx = p$. The CDF function calculates the following integral: $\int_{-\infty}^{x} f(x)dx$. If the r.v. is discrete, replace the integral signs by sum signs and the PDF by the PMF. The sampling function will draw a random realization from whichever distribution. All other parameters are the function parameters in the density / PMF that you should be familiar with from the class notes (and your cheat sheet).

Problem 1 This question is about midair collisions over American soil — an event which whose probabilities were first predicted using Bayesian methods as we learned about in McGrayne's book.



Since there are on the order of 10's of millions of flight hours per year and collisions are rare, this can be modeled using the Poisson random variable. Table 2 shows collision data in United States airspace by year from 1969 - 1978.

Year	Number of Collisions
1969	23
1970	32
1971	27
1972	24
1973	24
1974	32
1975	28
1976	30
1977	34
1978	33

Table 2: Official US Midair collision data for $n = 10$ years from 1969-1978

(a) [3 pt / 3 pts] The model we're using is $X_1, \ldots, X_n \overset{exch}{\sim} \text{Poisson}(\theta)$ where each X_i is the model for number of collisions in a given year. Is this a realistic model? Why or why not? There is no "correct" answer here but I expect you to defend whatever answer you write using the concepts we discussed in class.

No. Neither the # of flights / flight-hours nor the probability of crashes (which constitute θ) are the same year-to-year. Thus the yearly data should not be permuted – order matters and exchangeability is not realistic.

(b) [4 pt / 7 pts] Despite what you wrote in (a), assume the model is exchangeable for the rest of the problem. We are interested in inference for θ and we have no subjective prior opinion so we need to employ an objective prior for θ. However, θ will determine insurance rates and people's decision to fly, so we're really interested in some cost function $c(\theta)$ which will be monotonic but complicated. Which objective prior should we use to avoid getting different answers when new c functions are considered? Make sure you write $\theta \sim$ something below and make sure you specify the parameters as numbers.

Jeffrey's prior: $\theta \sim \text{Gamma}\left(\alpha = \frac{1}{2}, \beta = 0\right)$

(c) [5 pt / 12 pts] Given the data in Table 2 and the prior in (b), provide a 99% credible region for the true mean number of midair collisions from 1969-1978. Use the notation from Table 1 but do not solve numerically.

$X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\theta)$ & $\theta \sim \text{Gamma}(\alpha, \beta) \Rightarrow \theta | X \sim \text{Gamma}(\Sigma x_i + \alpha, n + \beta)$

$= \text{Gamma}(287.5, 10)$

$\Rightarrow CR_{\theta, 99\%} = \left[\text{qgamma}(.005, 287.5, 10), \text{qgamma}(.995, 287.5, 10)\right]$

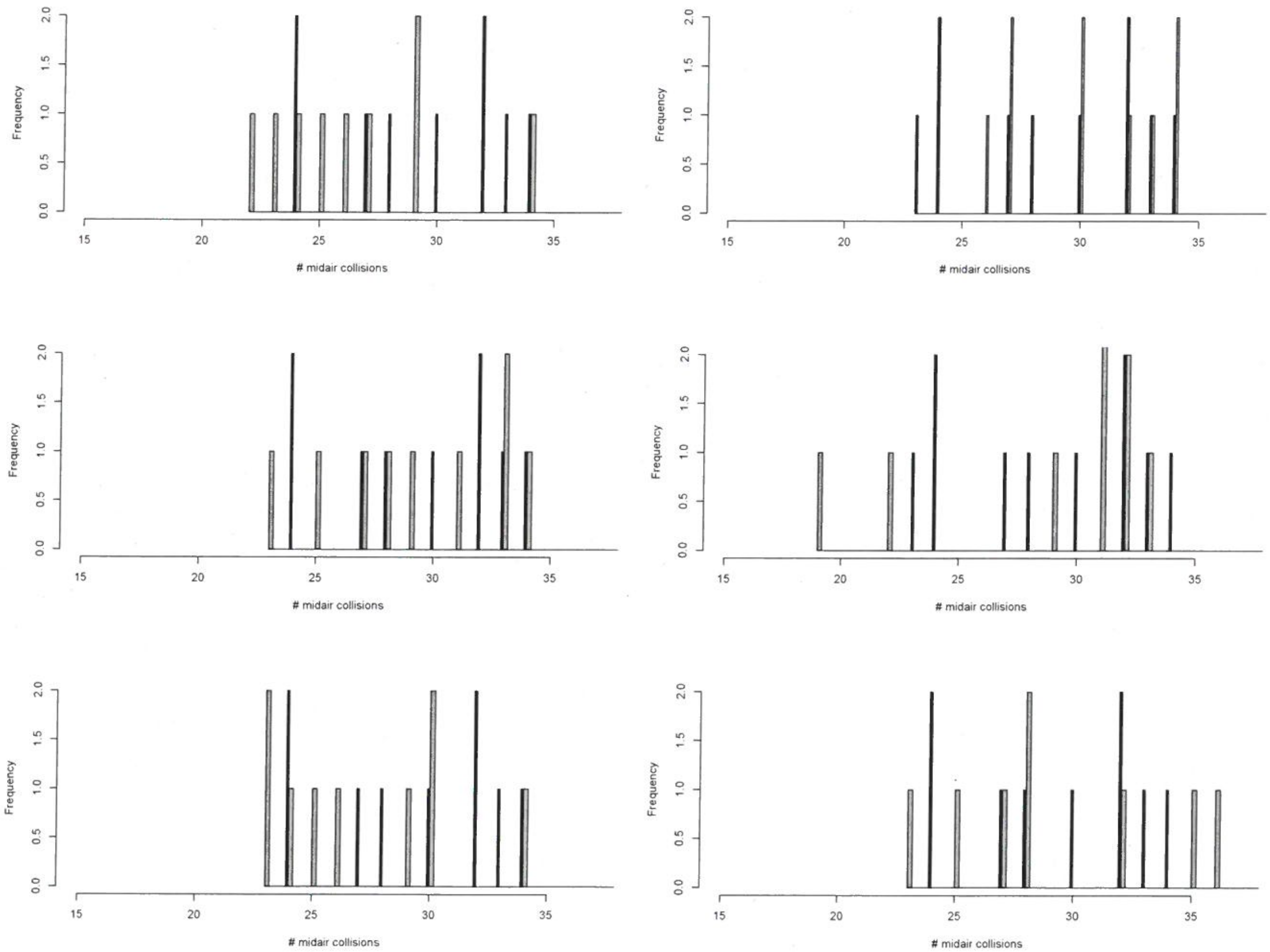
(d) [5 pt / 17 pts] Given the data in Table 2 and the prior in (b), calculate the probability that in 1979 we would see 36 or more midair collisions. Use the notation from Table 1 but do not solve numerically.

$X^* | X \sim \text{NegBin}\left(\Sigma x_i + \alpha, \frac{n+\beta}{n+\beta+1}\right) = \text{NegBin}(287.5, .909)$

$P(X^* \geq 36 | x) = 1 - P(X^* \leq 35 | X) = 1 - \text{pnbinom}(35, 287.5, .909)$

discrete support

(e) [3 pt / 20 pts] We would like to check if our model is valid using a graphical check. So we look at our dataset (which we denote x) against samples from $\mathbb{P}(X^* \mid X = x)$ where the dimension of X^* is the same as the dataset, n. Here are a few below. The real data is colored black and the posterior samples are in grey.



Does this model seem to be appropriate for the data? Why or why not?

Yes. The grey distribution is indistinguishable from the black (the true data) at least in 6 samples. We may be able to say more with more replication samples or by calculating some comparison test statistic (which we didn't cover in class).

(f) [3 pt / 23 pts] [Extra Credit] Exactly how did I generate those figures? For any one of the above six figures, write the steps necessary. (All the figures are identical modulo random sampling). Use the notation from Table 1.

Problem 2 We want to predict sale price of new cars based on fuel efficiency measured in miles per gallon (MPG).

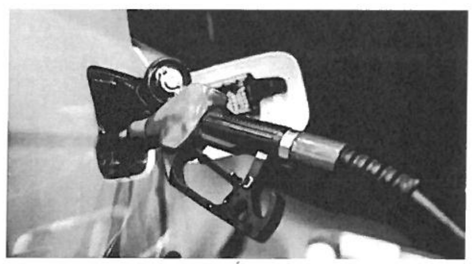


Table 3 displays sample data for a sample of 10 Japanese cars.

Fuel efficiency (nearest MPG)	Price New (nearest $1000)
31	23
31	29
30	27
29	25
30	28
30	27
28	28
32	29
29	26
28	27

Table 3: Fuel Efficiency and sale price for 10 Japanese cars

(a) [3 pt / 26 pts] If you were building a prediction model for this problem using data such as the data pictured in Table 3, what would the x variable be and what would the y variable be?

X: fuel efficiency in MPG

y: sale price in $1000's

(b) [3 pt / 29 pts] On a similar dataset of size $n = 10$, you calculate $\bar{x} = 30.08$, $\bar{y} = 29.01$, $r = 0.946$, $s_x = 0.609$ and $s_y = 0.534$. Use this information to determine the least squares estimates for the best fit line. Recall, $b_1 = r\frac{s_y}{s_x}$.

$b_1 = .946 \frac{.534}{.609} = .829$, $b_0 = \bar{y} - b_1\bar{x} = 29.01 - .829 \cdot 30.08 = 4.07$

(c) [3 pt / 32 pts] Is your answer in (b) dependent on assuming the OLS assumptions? Yes / no. Circle the answer.

(d) [3 pt / 35 pts] If you were to use a ridge penalty to determine your best fit line, would your intercept be smaller or larger? Circle the answer.

(e) [3 pt / 38 pts] If you were to use a ridge penalty to determine your best fit line, would your slope be smaller or larger? Circle the answer.

(f) [3 pt / 41 pts] Would you be able to use logistic regression to predict car price? Yes / no. Circle the answer.

(g) [6 pt / 47 pts] Let's say you have no prior information on where your intercept should be, but prior information on where your slope should be indexed by m such that $\beta_1 \sim \mathcal{N}(0, \sigma^2/m)$. Find the most parsimonious kernel for the posterior of $\mathbb{P}(\beta_1, \beta_0 \mid \boldsymbol{X}, y, \sigma^2)$ assuming the OLS assumptions.

$$P(\beta_1,\beta_0 | X, y, \sigma^2) \propto P(Y | \beta_0, \beta_1, \sigma^2, X)\, P(\beta_1, \beta_0 | \sigma^2, X)$$

$$= P(Y | \beta_0, \beta_1, \sigma^2, X)\, P(\beta_1 | \sigma^2, X)\, \underbrace{P(\beta_0 | \sigma^2, X)}_{1}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2} \frac{1}{\sqrt{2\pi\frac{\sigma^2}{m}}} e^{-\frac{m}{2\sigma^2}\beta_1^2} (1)$$

$$\propto e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + m\beta_1^2\right)}$$

(h) [3 pt / 50 pts] [Extra credit] Solve for this "half ridge" estimator (denote it $B_{\tilde{R}}$) proposed in (f). Hint: the ridge estimator is $B_R := \left(\boldsymbol{X}^T\boldsymbol{X} + mI_p\right)^{-1}\boldsymbol{X}^TY$.

(i) [3 pt / 53 pts] [Extra credit] If you were using a Bayesian model for β with the OLS assumptions and the ridge prior, how would you estimate $\mathbb{E}\left[||\beta||^2 \mid \boldsymbol{X}, y, \sigma^2\right]$? Hint: use the Monte Carlo / sampling concept we learned about in class and function(s) in Table 1.

(j) [3 pt / 56 pts] [Extra credit] Why does ridge regression frequently perform better than regular least squares regression on new observations?

Problem 3 This question is about building models for the prices of cars.

The 2016 Honda Accord sells at many different dealerships in New York City but sell it for more and some for less. We'll assume that the final negotiated price is distributed normally because it's most likely the sum of many different negotiation factors.

Our goal here is to determine the mean price at a certain car dealership in Astoria that people have been saying is "too cheap" and if it's too cheap, Honda corporate may wish to investigate.

Below are the sample average selling prices (in USD) of Honda Accords from 16 other car dealerships also in the NYC area that serve as a comparison:

```
22889.80 21159.16 23796.71 19132.65 23450.63 24088.28 19852.37 21306.45
24434.05 23150.34 21690.09 20640.79 21973.45 21984.48 22326.00 22239.98
```

(a) [4 pt / 60 pts] The average of the above prices is 22132.20 and the standard deviation is 1496.30. Create a prior distribution using empirical Bayes for the mean price at the Astoria car dealership. Assume we are using the normal-normal model.

$$\theta \sim N\left(\mu_0 = 22132.2,\ \tau^2 = 1496.3^2\right)$$

remember it's squared here

(b) [4 pt / 64 pts] Assume that each Accord's price at the Astoria dealership is normal and exchangeable. Is this a good model? Why or why not? There is no "correct" answer here but I expect you to defend whatever answer you write using the concepts we discussed in class.

No. The selling price may be dependent on availability/stock at the time, the specific salesman, etc. So θ & σ^2 may differ from X_i to X_j so exchangeability is not assumable.

(c) [4 pt / 68 pts] Despite what you wrote in (b), assume the model is exchangable for the rest of the problem. The nationwide average standard deviation for a Honda Accord selling price we're going to assume is $\sigma = \$1000$, an assumption we will relax later. Given a sample with average $\bar{x}$ and sample size n, what is the distribution of the mean price of a car from this shady Astoria dealership? Assume your prior from (a).

$$\theta \mid X, \sigma^2 \sim N\left(\frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}},\ \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} \right)$$

where $\sigma^2 = 1000^2$, $\mu_0 = 22132.2$, $\tau^2 = 1998.3^2$

(d) [4 pt / 72 pts] You and your colleague go down to the Astoria dealership undercover and ask to buy a Honda. After much negotiation, they will sell it to you for \$19,000 and they will sell it to your colleague for \$18,200 but they sense something suspicious so you hesitate to send another one of your guys down there to do another faux negotiation. Unfortunately, we're going to have to estimate the mean with just $x_1 = 19000$ and $x_2 = 18200$. What is your best guess of the mean price of Honda Accords sold here? Assume your prior from (a). Compute explicitly as a number rounded to two decimals.

$$\hat{\theta}_{mmse} = \hat{\theta}_{MAE} + \hat{\theta}_{map} = \frac{\frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma^2}} = \frac{\frac{19000 + 18200}{1000^2} + \frac{22132.2}{1996.3^2}}{\frac{2}{1000^2} + \frac{1}{1996.3^2}} = \frac{.0372 + .008085}{.000002} = \boxed{\$19244.70}$$

$\nearrow \Sigma x_i$

(e) [4 pt / 76 pts] What is the shrinkage value (which we have been denoting ρ) for this estimate? Compute explicitly as a number rounded to two decimals.

$$\rho = \frac{\sigma^2}{n\tau^2 + \sigma^2} = \frac{1000^2}{2(1996.3)^2 + 1000^2} = \boxed{.18}$$

(f) [6 pt / 82 pts] Based on this data, we wish to test if this dealership is selling Honda Accords below the manufacturer sugested retail price (MSRP) of \$22,205 — if so, they would be subject to a fine. Calculate a p-value for this test below by using notation from Table 1 but do not solve numerically.

$H_0: \theta \geq 22205$

$H_a: \theta < 22205$

$$SE[\theta|x] = \sqrt{\frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}} = \sqrt{\frac{1}{\underset{\substack{\uparrow \\ \text{from (d)}}}{.000002}}} = 639.31$$

$$p\text{val} = P(H_0|x) = P(\theta \geq 22205|x) = 1 - P(\theta < 22205|x) = 1 - \text{pnorm}(22205, \underset{\substack{\uparrow \\ \text{from (d)}}}{19244.7}, 639.31)$$

(g) [6 pt / 88 pts] What is the probability I get a really good deal — that I can buy a car from these Astoria people for under \$17,000? Use the notation from Table 1 but do not solve numerically.

$$X^*|X \sim N(\theta_p, \sigma_p^2 + \sigma^2) = N(19244.7, 639.31^2 + 1000^2) = N(19244.7, 1186.9^2)$$

be careful here

$$P(X^* < 17000|x) = \text{pnorm}(17000, 19244.7, 1186.9)$$

(h) [4 pt / 92 pts] If you were to estimate (g) without knowledge that $\sigma = \$1000$ but instead use a uninformative prior for σ^2, would the probability of getting the same really good deal be greater than, less than or equal to your answer in (g)? Explain why.

It would be greater than the answer in (g) since the resulting T or T-like mixture distribution will have fatter tails to account for the uncertainty in σ^2.

(i) [6 pt / 98 pts] We will continue to not rely on the nationwide average of $\sigma = \$1000$. Here, instead of an uninformative prior, use the data from the problem header to *estimate* a conjugate prior for σ^2. Pretend those 16 car dealerships are cars themselves. Be clear about everything your estimate relies on. Round the parameters to two decimal points.

The conjugate prior is $\sigma^2 \sim \text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$

We "see" $n_0 = 16$ and $\sigma_0^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(y_i-\theta)^2$ which is guesswork since we don't know θ but we can estimate it with s^2

$\approx s^2 = 1496.3^2$

$\Rightarrow \sigma^2 \sim \text{InvGamma}(8, 1.79\times 10^7)$

(j) [4 pt / 102 pts] We want to use the answer from (i) to fit a *conjugate* normal-normal model (with σ^2 unknown). This requires solving for (m) in the $\mathbb{P}(\theta \mid \sigma^2)$ prior. So we set s^2 from (a) equal to σ^2/m and solve for m and we get $m = 0.45$ rounded to the nearest two digits. What is our prior on θ, σ^2 now? You can notate your answer in terms of standard densities and you do not have to simplify it to a kernel.

$$P(\theta,\sigma^2) = P(\theta|\sigma^2)\,P(\sigma^2) = N\left(\mu_0, \frac{\sigma^2}{m}\right)\text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0\sigma_0^2}{2}\right)$$

$$= N\left(22132.2, \frac{\sigma^2}{0.45}\right)\text{InvGamma}\left(8, 1.79\times 10^7\right)$$

(k) [6 pt / 108 pts] Given the data in (d) which is $x_1 = 19000$ and $x_2 = 18200$, what is your best guess of the mean price of Honda Accords sold here? Assume your conjugate prior from (j).

$\bar{x} = \frac{19000+18200}{2} = 18600$

not clear for solution

$$\theta|X \sim T_{n_0+n}\left(\frac{n}{n+m}\bar{x} + \frac{m}{n+m}\mu_0, \_\_\_\right)$$

$\hat\theta_{MAE} = \hat\theta_{MAP} = \hat\theta_{MMSE}$

$$\Rightarrow \text{"}E[\theta|X]\text{"} = \frac{2}{2+0.45}18600 + \frac{0.45}{2+0.45}22{,}132.2 = \boxed{\$19{,}248.80}$$

(l) [4 pt / 112 pts] If you were to answer (k) but this time assume an independent prior for θ from (a) and independent prior for σ^2 from (i) and *not* use the conjugate prior in (k), you would not be able to simply compute an estimate of mean price. Explain one way in which you could go about estimating this mean now. Provide one sentence of explanation *only*. I am not looking for you to do any computation or describe a computer program.

No. This would be a semiconjugate model where you can use either grid sampling or a normal approximation to sample from $P(\sigma^2|X)$.