STA 360/601: Fall 2018, Midterm

October 11, 2018

Community Standard

To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

I have adhered to the Duke Community Standard in completing this exam.

Name: _____

NetID: _____

Signature: _____

Please write your name at the top of every page!

Common distributions

Normal with mean θ and variance σ^2: $p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(y-\theta)^2)$ for $y \in \mathbb{R}$

Multivariate (p-dimensional) normal with mean vector θ and covariance matrix Σ: $p(y|\theta,\sigma^2) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}}\exp(-\frac{1}{2}(y-\theta)^t\Sigma^{-1}(y-\theta))$ for $y \in \mathbb{R}^p$

Exponential with mean λ and variance λ^2: $p(y|\lambda) = \frac{1}{\lambda}\exp(-\frac{y}{\lambda})$ for $y > 0$

Gamma with mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$: $p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}\exp(-\beta y)$ for $y > 0$

Inverse Gamma with mean $\frac{\beta}{\alpha-1}$ and variance $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$: $p(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{-(\alpha+1)}\exp(-\frac{\beta}{y})$ for $y > 0$

Uniform distribution on $(0,1)$: $p(y) = 1$ for $y \in [0,1]$

Poisson distribution with mean θ: $p(y|\theta) = \frac{\exp(-\theta)\theta^y}{y!}$ for y a non-negative integer

Σ is an inverse Wishart distribution with parameters (ν_0, S_0^{-1}):

$p(\Sigma) \propto |\Sigma|^{-(\nu_0+p+1)/2}\exp(-\mathrm{tr}(S_0\Sigma^{-1})/2)$ and $E[\Sigma^{-1}] = \nu_0 S_0$

Beta distribution with mean $\frac{a}{a+b}$: $p(y|a,b) = \frac{1}{B(a,b)}y^{a-1}(1-y)^{b-1}$ for $y \in [0,1]$ and $a > 0, b > 0$

1. (2 points each) State/Define the following:

   (a) State Bayes' theorem.

   (b) For a model $p(y|\theta)$, what is Jeffrey's prior for θ?

   (c) Define infinite exchangeability.

   (d) TRUE or FALSE: The estimator that minimizes Bayes risk under quadratic loss is the posterior median.

   (e) TRUE or FALSE: Buffon's needle experiment can be used to approximate π with arbitrary precision.

2. (10 points) You are studying the length of different comments on internet discussions and following some sage advice you decide to think about them as being distributed as log-normal. That is let $X_1, \ldots, X_n$ be n observations of different comment lengths. Then conditional on θ and σ^2 they are log-normal with pdf given by $\frac{1}{x\sigma\sqrt{2\pi}} \exp(-\frac{(\ln x - \mu)^2}{2\sigma^2})$. The posterior mean of a log-normal variable with parameters μ and σ^2 is $\exp(\mu + \sigma^2/2)$.

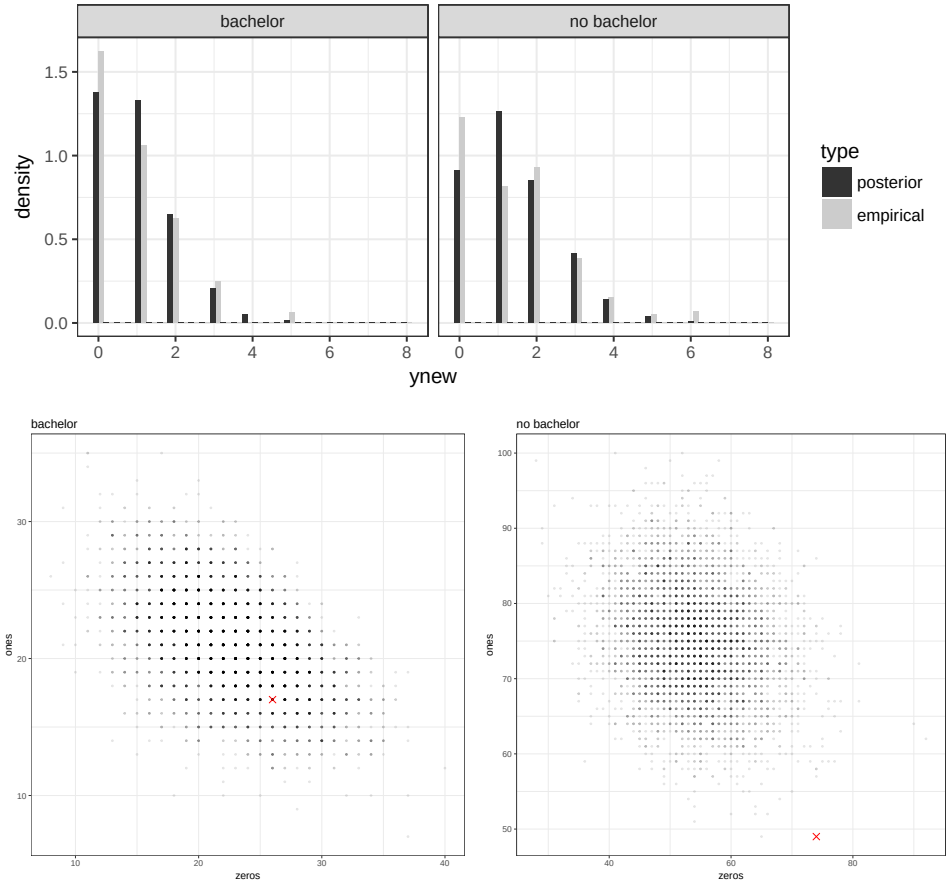
   (a) Show that if X is log-normal then $Y = \ln(X)$ is normally distributed.

   (b) Consider σ^2 to be fixed and known, find the maximum likelihood estimate of μ.

   (c) Consider σ^2 to be fixed and known, find the conjugate family of priors for the parameter μ when X is log-normal(μ, σ^2).

   (d) Write the posterior mean as a linear combination of the prior mean and the MLE. Give an interpretation to the parameters in the conjugate prior.

3. (10 points) Posterior predictive checks: Let θ_A and θ_B be the average number of children of men in their 30s with and without a bachelor's degrees, respectively. In the sample there are 58 men with a bachelor's degree and 218 men without one. The two samples are modeled as independent Poisson variables with Gamma priors. We plot the posterior predictive distribution of the data for the two models as well as the posterior predictive distribution of the number of zeros versus the number of ones in the two models (note that the observed value is marked by an "X").



(a) What do the above pictures tell you about the models? Does the Poisson model capture everything in the data well? What can we learn from the collected data? Comment on whether you trust analysis about θ_A and θ_B (separately) based on this? How about comparisons of θ_A to θ_B?

(b) Write out the scheme for getting the data to draw the posterior predictive distributions above. For each step in the scheme, state which distribution you are sampling from.

4. (10 points) Normal models. Make sure to plug in the numbers in the problem!

(a) Let $X_1, \ldots, X_{10}|\theta$ be iid samples from $N(\theta, 1)$ (variance= 1). Let the sample mean $\bar{X} = 3$. Assume the prior is $\theta \sim N(0, 5)$. Compute the posterior distribution of θ.

(b) Let $Y_1, \ldots, Y_{10}|\theta$ be another sample that is iid from $N(\theta, 2)$ (variance= 2) and is also independent of the first sample conditional on θ. The sample mean is $\bar{Y} = 2$. Compute the posterior distribution of θ using both samples (Xs and Ys).

(c) You have now gotten a third sample $Z_1, \ldots, Z_{10}|\theta$ iid $N(\theta, 3)$ (variance= 3). Let the measured sample mean be $\bar{Z} = 1$. This sample is yet again independent of the first two samples, but unfortunately your measuring device is not great and truncates observations greater than 2 down to 2. The last 2 observations (Z_9, Z_{10}) in this sample were truncated (so you know that $Z_9 \geq 2, Z_{10} \geq 2$ — this also means that $\bar{Z} = (\sum_{i=1}^{8} Z_i + 2 + 2)/10$). Compute posterior distribution based on all three samples (the posterior here is not a standard distribution so write what it is proportional to).

(Hint: To simplify the calculations, figure out how to plug in the result from (a) into part (b) and the result from (b) into part (c).)

5. (10 points) Researchers are interested in estimating the frequency of the homozygous rate variant of a particular single nucleotide polymorphism in the TP53 gene. In a random sample, out of 173 individuals, 6 had the homozygous rare variant. Let π denote the proportion of individuals with the homozygous rare variant in the general population.

   (a) Let each individual's homozygous rare variant status is denoted by X_i, what is a reasonable model for such data? What assumptions do you need for it to hold?

   (b) What is the maximum likliehood estimate of π?

   (c) Using a uniform prior distribution for π, find the posterior distribution for π.

   (d) Write the posterior mean as a linear combination of the prior mean and the MLE.

   (e) (Bonus, 2 points) The homozygous rare variant may be denoted by aa. If the frequency of the rare allele a is π_a in the population then the frequency of the homozygous rare genotype is related to the allele frequency by $\pi = \pi_a \pi_a$. Using this relationship and your posterior distribution above, find the posterior distribution of the rare allele frequency π_a.