# MATH 640: Exam 1

Name:______________________________________________

## Instructions

1. Do not discuss this midterm with anyone other than Professor Meyer or the TA; and you may only ask clarifying questions.
2. Your responses to this exam must be typed.
3. This cover sheet *must* be the first page of your submission.
4. If you wish to cite a result or derivation from lecture, you may do so but be sure it is clearly cited (list the slide from the Note Set or the example) and relevant.
5. Please submit the exam to the assignment page on Canvas by **11:59pm on Sunday, March 14**.
6. Late submissions will be accepted up to 24 hours after the deadline, however they will be penalized: 4 points off for every six hours the submission is late.

| Portion | Question | Points | Score |
|---|---|---|---|
| Theory | 1 | 15 | |
| | 2 | 15 | |
| | 3 | 20 | |
| | 4 | 10 | |
| Computing | 1 | 20 | |
| | 2 | 20 | |
| Total | | 100 | |

# Theory

1. Suppose you wish to model a random sample of standard deviations, which you believe to be small, using the half-Normal distributions. The density is then

$$f(s_i) = \left(\frac{2}{\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2\sigma^2}s_i^2\right)1(s_i > 0),$$

for $\sigma^2 > 0$. Use this to answer the following.

   (a) (5 points) Let $s_i$ be an iid sample, $i = 1, \dots, n$, from a half-Normal and define the transformation $\tau^2 = 1/\sigma^2$, the precision. State the likelihood in terms of the precision and determine Jeffrey's prior for $\tau^2$.

   (b) (5 points) Find the posterior distribution that results from using the prior you found in (a).

   (c) (5 points) Explain why placing placing the non-informative prior on the precision, $\tau^2$, is equivalent to placing a non-informative prior on the variance, $\sigma^2$.

2. The log-normal distribution is useful for modeling skewed data and has the density

$$f(x_i) = \frac{1}{x_i\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(\log\{x_i\} - \mu\right)^2\right],$$

for $\sigma^2 > 0, \mu \in \mathbb{R}$ and $x_i > 0$. The joint non-informative prior for $\mu$ and $\sigma^2$ is $\pi(\mu, \sigma^2) \propto (\sigma^2)^{-1}$. Use this to answer the following.

   (a) (5 points) Assuming you have an iid sample of $x_i$'s, find the conditional posterior distribution of $\mu|\sigma^2$. (Hint: it may be useful to define $\bar{x}_\ell$ as $\frac{1}{n}\sum_{i=1}^n \log\{x_i\}$, i.e. the log of the geometric mean.)

   (b) (10 points) Now derive the marginal posterior distribution for $\sigma^2$.

3. The Gompertz distribution can be used to analyze survival times. It is parameterized by two parameters: a scale parameter $b$ and a shape parameter $\eta$. Its density is

$$f(t_i) = b\eta e^{bt_i+\eta} \exp\left(-\eta e^{bt_i}\right)$$

for $b, \eta > 0$ and $t_i \in [0, \infty)$. Assume that $b$ is fixed at $b_0$, but $\eta$ is unknown. Further assume that the prior on $\eta$ is $\pi(\eta) \propto \eta^{-1}$. Use this to answer the following.

   (a) (5 points) Suppose we have an iid sample of $n$ survival times that we assume to be Gompertz. Determine the likelihood and state the posterior.

   (b) (5 points) Find the posterior mode.

   (c) (5 points) Find the observed information.

   (d) (5 points) Using your answers in (b) and (c), develop a strategy for sampling from the posterior distribution of $\eta|t_1, \dots, t_n, b_0$.

4. (10 points) Suppose we are running a Phase I Clinical trial to determine toxicity of a new drug. There are four possible responses to the drug: no toxic event, mildly toxic event, moderately toxic event, and severely toxic event. As a stopping rule, we've determined that after $x_0 = 3$ severely toxic events, we will stop the trial. This process can then be modeled using a negative multinomial which has as its mass function

$$P(X_0 = x_0, X_1 = x_1, X_2 = x_2, X_3 = x_3) = \Gamma\left(\sum_{i=0}^{3} x_i\right)\frac{\theta_0^{x_0}}{\Gamma(x_0)}\prod_{i=1}^{3}\frac{\theta_i^{x_i}}{x_i!},$$

where the "failure" count is denoted by $X_0$. In this case a "failure" is having a severely toxic event. We conduct the trail and reach our stopping rule, giving us a single vector of counts with the first element defined by the stopping rule and the remaining elements determined by the observed counts in the other categories. Determine the likelihood and find a conjugate prior for the vector of probabilities, $\boldsymbol{\theta} = [\ \theta_0 \quad \theta_1 \quad \theta_2 \quad \theta_3\ ]$. Then, using your conjugate prior, suggest a non-informative prior.

# Computing

(a) Wind speed measurements were taken from New York's LaGuardia Airport over the course of 153 days starting in May and ending in September. We are interested in building a Bayesian model to examine attributes of the distribution of wind speeds. One common way to model wind speeds is using the Rayleigh distribution which is parameterized by its mode, $\theta > 0$ and has the density:

$$f(w_i) = \frac{w_i}{\theta^2} \exp\left(-\frac{w_i^2}{2\theta^2}\right).$$

With the distribution of the mode, we can empirically estimate the distribution of the mean $\left(\text{calculated as } \theta\sqrt{\pi/2}\right)$ as well as the median $\left(\text{calculated as } \theta\sqrt{2\log(2)}\right)$.

Assuming the sample is iid, first determine the likelihood and then find a conjugate prior for $\theta^2$, making sure to state the resulting posterior. For your hyper-parameters, select them to make the prior non-informative—justify your choice. Using the data in the file `wind.txt`[1] (wind speeds in mph), find the posterior distribution of the mode, mean, and median wind speeds at LaGuardia. Then find the predictive probability that the wind exceeds 15 mph on a day between May and September in 2021. Calculate summary statistics for each measures along with credible intervals. Generate $B = 20000$ samples and set the seed to 6302.

(b) Data from the Capital Bikeshare system was collected from 2011 and 2012 along with a set of predictors with the aim of building a model to assess the impact of different phenomenon, weather and workday related, on daily system usage. The dataset is in the file `day.txt` and contains the following variables:

i. `casual`: count of casual users on a day (outcome # 1)
ii. `registered`: count of registered users on a day (outcome # 2)
iii. `yr`: year (coded 0 for 2011, 1 for 2012)
iv. `holiday`: whether day is a holiday or not
v. `workingday`: 1 if day is neither weekend nor holiday, 0 otherwise
vi. `temp`: normalized temperature in Celsius
vii. `atemp`: normalized "feeling" temperature in Celsius
viii. `hum`: normalized humidity, the values were divided by 100
ix. `windspeed`: normalized wind speed, the values were divided by 67

The rows of the dataset correspond to days.

Build two regression models, one with casual users as the outcome, the other with registered users. For each model, determine the "best" set of predictors supporting your choice with Bayesian inference results. Before modeling, check each outcome variable to ensure normality is a reasonable assumption, transform accordingly if not. Take $B = 20000$ samples for each, setting the seed to 97 for the casual users model and 726 for the registered users. Compare and contrast the final models discussing any differences, if any, and suggest possible reasons for what you see. Also discuss any similarities you see between the models. Provide relevant posterior summaries for all model parameters.

Finally, Leap Day 2012 (February 29, 2012) was withheld from the dataset. Using its covariates, predict both the number of casual and registered users: `yr` = 1, `holiday` = 0, `workingday` = 1, `temp` = 0.344348, `atemp` = 0.34847, `hum` = 0.804783, and `windspeed` = 0.179117. Be sure to provide credible intervals for both of your predicted values. Set the seed to 2011 for your casual-model prediction and 2012 for your registered-model prediction.

[1] The assumption of independence in time series data is usually suspect, however, the autocorrelation among the samples is quite low, even for successive days and drops off quickly suggesting that we can feel OK with this assumption.