36-724 Homework 1
Due: Monday, January 30, 14:30 EST, by email to Dansci (724homeworksgohere@gmail.com).

Please submit your raw R code along with your solution set as a PDF (in LaTeX, scanned-in pencil/pen, whichever for now). Name your files in the fashion of yourname_hw1.pdf and yourname_hw1.R.

1. Bayes' Theorem for discrete events, and avoiding the prosecutor’s fallacy.
   Let $T$ be an event of a positive test result for a disease, and $D$ be the event of actually having the disease. The test for the disease is labelled at being 99% accurate, meaning that the false positive and negative probabilities are both 0.01. The incidence of the disease in the general population is one in fifty thousand.
   
   (a) Using Bayes' Theorem, calculate the probability that a person from the general population who tests positive actually carries the disease.
   
   (b) Confirm this result by simulation in R:

   ```R
   #simulate the disease status for one million individuals
   people <- 1000000
   disease <- rbinom(people, 1, 1/50000)
   positives <- sum(disease)
   positive.test <- rep(NA, people)
   positive.test[disease==1] <- rbinom(positives, 1, 1-0.01)
   positive.test[disease==0] <- rbinom(people-positives, 1, 0.01)
   outcomes <- 2*positive.test + disease
   #(true pos: outcomes=3; true neg, 0; false pos, 2; false neg, 1)
   c(sum(outcomes==3), sum(outcomes==0), sum(outcomes==2), sum(outcomes==1))
   ``

   Run this code 10 times and have the program report the number of true and false, negatives and positives, as well as the fraction of true positives over all positives.

2. Generating random variates through transformation.
   (a) Show by density transformation that if $U \sim Unif(0, 1)$, then $Y = U^{1/\alpha} \sim Beta(\alpha, 1)$.
   
   (b) Verify this result by simulation in R for $\alpha = 5$ by comparing simulations of $U^{1/5}$ to those from $Y \sim Beta(5, 1)$. You may use a Q-Q plot or statistical test; have the R program output this result.
   
   (c) Show by CDF transformation that $U \sim Unif(0, 1)$, then $Y = U^{-1/\alpha}$ is Pareto distributed with shape $\alpha$ and scale 1. (You can now generate your own Pareto variates, a function not included by default in R!)

3. Continuous mixture models.
   (a) Show by integration that if $X|V \sim N(0, 1/V)$ and if $V \sim Gamma(shape = a/2, rate = a/2)$, then marginally $X$ has a Student-t distribution with degrees of freedom $a$. 

1
(b) Verify this by simulation in R for \( n = 10000 \) and \( a = 5 \), first by simulating the gammas (\( \text{gams} \leftarrow \text{rgamma}(nn, \text{shape}=aa/2, \text{rate}=aa/2) \)) and then the normals (\( \text{draws} \leftarrow \text{rnorm}(nn, 0, \text{sqrt}(1/\text{gams})) \)). Compare this to a t draw with \( \text{t.draw} \leftarrow \text{rt}(nn, aa) \).

(c) Show by integration that if \( Y|\lambda \sim \text{Poisson}(\lambda) \) and \( \lambda \sim \Gamma(\text{shape} = a, \text{rate} = b) \), than marginally \( Y \) has a negative binomial distribution (with support on the whole numbers). Express the parameters \((r, p)\) in terms of \( a \) and \( b \). Note that there are 4 different forms for the Negative Binomial – to which does this correspond?

(d) Verify this by simulation in R for \( n = 10000 \), \( a = 6 \) and \( b = 2 \), first by simulating the gammas (\( \text{gams} \leftarrow \text{rgamma}(nn, \text{shape}=aa, \text{rate}=bb) \)) and then the Poissons (\( \text{draws} \leftarrow \text{rpois}(nn, \text{gams}) \)). Compare this to a negative binomial draw with \( \text{nb.draw} \leftarrow \text{rnbinom}(nn, \text{trial.size, prob}) \) once you have solved for the two parameters.

4. Conjugate prior specifications.

(a) Show that the “rate” parameter in the Gamma distribution \((\beta, \text{for } p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x})\) has a Gamma conjugate prior. Determine the posterior parameters if the prior is distributed as Gamma(5, 10).

(b) Show that the size parameter \( \theta \) in the Uniform(0, \( \theta \)) has a Pareto conjugate prior. Note that the density of a bounded distribution is often best described by indicator functions, such as

\[
Pr(a < U < b) = \int_a^b \frac{1}{\theta} \mathbb{1}(u > 0) \mathbb{1}(u < \theta) du.
\]

5. Posterior inferences. Six hundred people are surveyed at random (from an infinite population) for their music preferences. In this survey, 320 prefer John Coltrane while 280 prefer Sonny Rollins. Let \( \phi \) be the fraction of people in the (extremely large) population who prefer John Coltrane, on which we want to make a posterior inference.

Assuming a prior uniform distribution on \( \phi \), determine the form of the posterior probability distribution \( p(\phi|Y) \), and using this, determine:

(a) The maximum a posteriori (MAP) density, or the most likely value for the fraction of Coltrane fans in the population – \( \arg \max_\phi p(\phi|Y) \). Determine this analytically.

(b) The MAP density, as determined numerically using the appropriate density (pdf) function in R.

(c) The posterior probability that more people prefer John Coltrane given this survey – that is \( Pr(\phi > 0.5|Y) \). Determine this numerically in R.

(d) The posterior probability that more than 51% of people prefer John Coltrane – \( Pr(\phi > 0.51|Y) \). Determine this numerically in R.

6. Rejection sampling. Consider a mixture distribution: with probability 1/3 we draw from a \( \text{Beta}(6, 2) \), and with probability 2/3 we draw from a \( \text{Beta}(2, 6) \).

(a) Graph the density of this mixture distribution by adding two \text{dbeta} terms over a sequence of values between 0 and 1.
(b) Write a rejection sampler for this distribution by using a Uniform density as a proposal distribution. Use your answer from the previous question to estimate a minimum value of the multiplier $M$, or derive it directly from the PDF of the mixture.

(c) Estimate the efficiency of this sampler (number of accepted draws over number of proposals) through simulation.

(d) Use the simulations from the previous step to compare to direct simulations from the mixture:

```r
first.comp <- rbinom(1, nn, 1/3);
draws <- c(rbeta(first.comp, 6, 2), rbeta(nn-first.comp, 2, 6))
```

Give a graphical comparison of the two simulations. Note that you can use the function `length(first.comp)` to find the length of the vector of draws.

7. Rejection sampling II. Consider drawing from the standard normal $(-2, 1)$ which is truncated below at 0. If we use an exponential proposal distribution with rate $\theta$, solve analytically for the choice of rate $\theta$ that maximizes the efficiency of the sampler. Demonstrate this in R by running the rejection sampler for 100,000 proposals each for values of $\theta$ in the neighbourhood of the maximum and showing that this target tends to outperform others.