Homework 1

Due Wednesday January 31st by end of day, via Canvas (typeset or legible scanned copy)

PUBH 8442: Bayes Decision Theory and Data Analysis

1. *Identifying Spam Text Messages*
   Bayes' rule can be used to identify and filter spam emails and text messages. The exercises below refer to a large collection of real SMS text messages from participating cellphone users.[1] In this collection, 747 of the 5574 total messages (13.40%) are identified as spam.

   a.) The word "free" is contained in 4.75% of all messages, and 3.57% of all messages both contain the word "free" and are marked as spam. What is the probability that a message contains the word "free", given that it is spam?

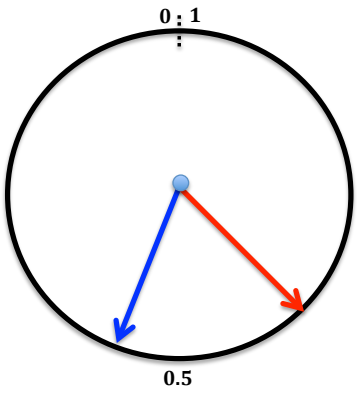
   b.) What is the probability that a message is spam, given that it contains the word "free"?

   c.) The word "text" (or "txt") is contained in 7.01% of all messages, and in 38.55% of all spam messages. What is the probability that a message is spam, given that it contains the word "text" (or "txt")?

   d.) Of all spam messages, 17.00% contain both the word "free" and the word "text" (or "txt"). For example, "Congrats!! You are selected to receive a free camera phone, txt ******* to claim your prize." Of all non-spam messages, 0.06% contain both the word "free" and the word "text" (or "txt"). Given that a message contains both the word "free" and the word "text" (or "txt"), what is the probability that it is spam?

   e.) Given that a message contains the word "free" but does NOT contain the word "text" (or "txt"), what is the probability that it is spam?

2. *Uniform spinner positions*
   A number line from 0 to 1 is attached at its ends to form a circle. Two spinning arrows are placed in the middle of this circle (see below). The blue arrow is spun once, so that it is pointing to a random position on the circumference of the circle, unknown to you. Then, the red arrow is spun, and you are told only whether it lands on a number that is greater or less than the blue arrow; this is repeated several times while the blue arrow remains fixed.



[1] Almeida, T., Hidalgo, J., Yamakami, A., "Contributions to the Study of SMS Spam Filtering: New Collection and Results," *Proceedings of the 2011 ACM Symposium on Document Engineering (DOCENG'11)*, Association for Computing Machinery, Mountain View, CA, USA, 2011.

a.) What is the probability density for the position of the blue arrow, given that the red arrow has landed on a greater value H times and a lower value L times?

b.) Give a point estimate for the position of the blue arrow, given that the red arrow has landed on a greater value H times and a lower value L times.

3. Derive the distribution for the posterior mean for the normal-normal model, with multiple observations, given on slide 16 of the *Prior and posterior* slide set.

4. Derive the prior and posterior predictive distributions for the normal-normal model, given on slide 22 of the *Prior and posterior* slide set.

5. Show that the Jeffreys prior based on the binomial likelihood is given by a $\text{Beta}(0.5, 0.5)$ distribution.