MATH 347 Homework 1 (Total 30 points)

Due: Thursday 9/12, at the beginning of the class

Name: _____

- **Print out this cover page and staple with your homework.**
- **Show all work. Incomplete solutions will be given no credit.**
- **You may prepare either hand-written or typed solutions, but make sure that they are legible. Answers that cannot be read will be given no credit.**
- **R graphical outputs must be printed instead of hand-drawn.**

1. (*4 points; 2 points each part*)
   It is estimated that roughly 8% of incoming email is spam. A spam filter has an accuracy rate of 92% for spam emails, and it incorrectly categorizes 3% as non-spam emails as spam.
   (a) What percent of all email is marked as spam?
   (b) If an email is marked as spam, what is the probability that it is indeed a spam email?

2. (*8 points; 2 points each part*)
   Hoff 2.1 (page 225)

3. (*4 points*)
   Let $X \sim \mathrm{N}(\mu, \sigma^2)$, what distribution does $Y = (X-\mu)/\sigma$ have? (Hint: review the change of variable material.)

4. (*3 points*)
   Let $X_1, \cdots, X_n \overset{iid}{\sim} \text{Normal}(\mu_X, \sigma_X)$, and $Y_1, \cdots, Y_m \overset{iid}{\sim} \text{Normal}(\mu_Y, \sigma_Y)$, and assume all X's and Y's are independent. Write out the joint density:
   $$f(X_1 = x_1, \cdots, X_n = x_n, Y_1 = y_1, \cdots, Y_m = y_m \mid \mu_X, \mu_Y, \sigma_X, \sigma_Y).$$

5. (*5 points*)
   We say a random variable X has a logistic distribution, if its cdf is
   $$F(x) = \frac{e^x}{1+e^x}, \quad -\infty < x < \infty.$$
   To compare this distribution with the standard normal distribution, use R to plot their pdf's (in the same graph). Suppose you draw 1000 random samples from each distribution, then which distribution will you expect to see more samples with extreme values (i.e., very large in absolute value, either positive or negative)? (Hint: in R, `dnorm` is the function to evaluate the pdf of a normal distribution, and `dlogis` is the function to evaluate the pdf of a logistic distribution.)

6. (*6 points; 2 points each part*)
   In the all-nighters example in class, we took a sample of $n = 10$ Vassar students and obtained $y = 3$ out of 10 who had pulled an all-nighter in last academic year. Recall that we came up with a prior distribution for the percentage p of all Vassar students who had pulled an all-nighter in last academic year as (in R):

```
priorvalues <- c(0, .1, .2, .3, .4, .5, .6, .7, .8, .9, 1)

priorprob <- c(1/23, 1/23, 7/23, 7/23, 3/23, 3/23, 1/23, 0/23, 0/23, 0/23, 0/23)
```

We used the pre-written RMarkdown file (on Moodle) to calculate the posterior probabilities of 11 values of p and plot the prior and posterior on the same graph.

Now consider another sample of $n_2 = 10$ with $y_2 = 5$ (use new notation for the previous sample $n_1 = 10$, $y_1 = 3$). Use the R script and make changes to perform the following:

(a) Treat two samples as one sample, that is, $n = n_1 + n_2 = 10 + 10 = 20$ and $y = y_1 + y_2 = 3 + 5 = 8$. Calculate the posterior probabilities of p given this new combined sample (n, y).

(b) Recall that a sequential update is to use the previous posterior from the first sample (n_1, y_1) as the prior for the second sample (n_2, y_2). Do a sequential update and check whether the posterior probabilities match with what you calculated in (a).

(c) Would sequential update make sense in this case? Why or why not?