STA 410/2102 — Practice questions for the second test

1. Suppose we're interested in the mean weight of cookies that are consumed before seminars in the Statistics department. To assist this investigation, the department has acquired two scales. One scale is fairly accurate, but the other is less accurate. Fortunately, both scales are unbiased. Both the errors of the scales and the weights of the cookies are normally distributed. The error made by a scale is unrelated to the weight of the cookie being weighed.

   Before each seminar, volunteers weigh the cookies, and record the weight they measure for each. Unfortunately, their main interest is in eating the cookies, so they don't bother to record which scale was used to weight each cookie. However, we do know that they have no preference for one scale over the other, and hence are equally likely to choose either one when weighing a cookie. From the list of cookie weights obtained in this way, we would like to estimate the mean weight of cookies from the department's cookie source.

   Write an R program that takes the list of measured cookie weights as an argument, and returns the maximum likelihood estimate for the mean cookie weight, found using an EM algorithm.

   Note that the model to use will have three parameters — μ, the mean weight of the cookies, σ_1, the standard deviation of weight measurements made by one of the scales, and σ_2, the standard deviation of weight measurements made by the other scale. The "missing data" for the EM algorithm will be the identifiers of which scale was used to measure each cookie.

2. Suppose we numerically evaluate the integral

$$\int_0^1 x^4 dx$$

   using the midpoint rule. Using 100 points, the approximation we get is 0.199983333625. Using 1000 points, the approximation we get is 0.199999833333363. The exact answer is of course 1/5. Estimate what approximation we will get if we use the midpoint rule with 2000 points.

3. Suppose we have i.i.d. data points $a_1, \ldots, a_n$ that are measurements of angles in radians, in the range of 0 to 2π. We decide to model this data with a form of the "von Mises" distribution that assigns probability density $K \exp(\cos(a_i - \theta))$ to data point a_i, where $K = 0.1257\ldots$, and θ is an unknown model parameter. Suppose that our prior distribution for θ is uniform over the range 0 to 2π. Given data $a_1, \ldots, a_n$, we wish to compute the Bayes factor for this von Mises model versus the simple model (with no parameters) that says the a_i are uniformly distributed over the range 0 to 2π. Recall that the Bayes factor for model A versus model B is the ratio of the probability of the data under model A to the probability of the data under model B, integrating over the parameters (if any) of each model with respect to the prior.

   Write an R function (or set of functions) that will compute this Bayes factor using the trapezoidal integration rule. The main function should take as arguments a data vector a and the number of points to use for the trapezoidal rule, and return the Bayes factor.