STAT 206 Final: Due Wednesday, December 12, 5PM

General instructions: Final must be completed as an R Markdown file. Be sure to include your name in the file. Give the commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used. **The final exam is open book/internet access, but absolutely no communicating with other humans.** Any questions you have must be directed to me.

Part I - Rescaled Epanechnikov kernel

The rescaled Epanechnikov kernel is a symmetric density function given by

$$f(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{for } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

1. Check that the above formula is indeed a density function.
2. Produce a plot of this density function. Set the X-axis limits to be -2 and 2. Label your axes properly and give a title to your plot.
3. Devroye and Gy"orfi give the following algorithm for simulation from this distribution. Generate iid random variables $U_1, U_2, U_3 \sim U(-1, 1)$. If $|U_3| \geq |U_2|$ and $|U_3| \geq |U_1|$, deliver U_2, otherwise deliver U_3. Write a program that implements this algorithm in R. Using your program, generate 1000 values from this distribution. Display a histogram of these values.
4. Construct kernel density estimates from your 1000 generated values using the Gaussian and Epanechnikov kernels. How do these compare to the true density?

Part II - Pine needles

In the article "Pine needle as sensors of atmospheric pollution", the authors use neutron-activity analysis to determine pollution levels, by measuring the Bromine concentration in pine needles. The investigators collect 18 pine needles from a plant near an oil-fired steam plant and 22 near a cleaner site. The data can be found at [http://faculty.ucr.edu/~jflegal/206/pine_needles.txt].

5. Describe the data using plots and summary statistics. Show that the data are **not** normally distributed by drawing an appropriate graphical display for each sample.
6. Take a log transformation of the values in each sample. Does it seems reasonable that the transformed samples are each drawn from a normal distribution? Test this formally using an appropriate test (of your choosing).
7. Now suppose that the authors of this study want to calculate an interval for the difference between the median concentrations at the two sites, on the original measurement scale. Write code to calculate a 95% bootstrap interval for the difference in the medians between the two samples. Summarize your conclusion in words.

Part III - Markov chains

Suppose you have a game where the probability of winning on your first hand is 48%; each time you win, that probability goes up by one percentage point for the next game (to a maximum of 100%, where it must stay), and each time you lose, it goes back down to 48%. Assume you cannot go bust and that the size of your wager is a constant $100.

8. Is this a fair game? Simulate ten thousand sequential hands to determine the size of your return. Then repeat this simulation 99 more times to get a range of values to calculate the expectation.
9. Repeat this process but change the starting probability to a new value within 2% either way. Get the expected return after 100 repetitions. Keep exploring until you have a return value that is as fair as you can make it. Can you do this automatically?
10. Repeat again, keeping the initial probability at 48%, but this time change the probability increment to a value different from 1%. Get the expected return after 100 repetitions. Keep changing this value until you have a return value that is as fair as you can make it.