STAT 206 Final

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 - Markov chains and fair games

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.

1. Is this a fair game? Simulate one hundred 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.
2. 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?
3. 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.

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].

1. 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.
2. 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).
3. 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 - Permutation tests

The Cram'er von Mises statistic estimates the integrated square distance between distributions. It can be computed using the following formula

$$W = \frac{mn}{(m+n)^2}\left[\sum_{i=1}^{n}(F_n(x_i) - G_m(x_i))^2 + \sum_{j=1}^{m}(F_n(y_j) - G_m(y_j))^2\right]$$

where F_n and G_m are the corresponding empirical cdfs.

1. Implement the two sample Cram'er von Mises test for equal distributions as a permutation test. Apply it to the `chickwts` data comparing the `casein` and `linseed` diets.

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 - Metropolis-Hasting algorithm

Suppose $f \sim \Gamma(2, 1)$.

1. Write an independence MH sampler with $g \sim \Gamma(2, \theta)$.
2. What is $R(x_t, X^*)$ for this sampler?
3. Generate 10000 draws from f with $\theta \in \{1/2, 1, 2\}$.
4. Write a random walk MH sampler with $h \sim N(0, \sigma^2)$.
5. What is $R(x_t, X^*)$ for this sampler?
6. Generate 10000 draws from f with $\sigma \in \{.2, 1, 5\}$.
7. In general, do you prefer an independence chain or a random walk MH sampler? Why?
8. Implement the fixed-width stopping rule for you preferred chain.

Part II - Beverton-Holt model

The dataset at [http://www.faculty.ucr.edu/~jflegal/fish.txt] contains 40 annual counts of the numbers of spawners `S` and recruits `R` in a salmon population. The units are thousands of fish. Spawners are fish that are laying eggs. Spawners die after laying eggs. Recruits are fish that enter the catchable population.

The classic **Beverton-Holt** model for the relationship between spawners and recruits is

$$R = \frac{1}{\beta_1 + \beta_2/S}, \quad \beta_1 > 0, \beta_2 > 0$$

where R and S are the number of recruits and spawners respectively.

Consider the problem of maintaining a sustainable fishery. The total population abundance will only stabilize if $R = S$. The total population will decline if fewer recruits are produced than the number of spawners who died producing them. If too many recruits are produced, the population will also decline eventually because there is not enough food for them all. Thus, only a balanced level of recruits can be sustained indefinitely in a stable population. This stable population level is the point where the 45° line intersects the curve relating R and S. In other words, it is the N such that

$$N = \frac{1}{\beta_1 + \beta_2/N}.$$

Solving for N we see that the stable population level is $N = (1 - \beta_2)/\beta_1$.

1. Make a scatterplot of the data and overlay the Beverton-Holt curve for a couple different choices of β_1 and β_2.
2. The Beverton-Holt model can be found by transforming $R \mapsto (1/R)$ and and $S \mapsto (1/S)$. That is,

   $$(1/R) = \beta_1 + \beta_2(1/S).$$

   This is a linear model with response variable $(1/R)$ and covariate $(1/S)$. Use least squares regression to fit this model to the fish dataset.
3. Find an estimate for the stable population level, where $R = S$ in the Beverton-Holt model.
4. Use the bootstrap to obtain the sampling distribution and standard error for the stable population level. Use the bootstrap to construct a 95% confidence interval for the stable population level.

Part III - Snowfall accumulations

The data set `buffalo` at [http://www.faculty.ucr.edu/~jflegal/buffalo.txt] contains annual snowfall accumulations in Buffalo, NY from 1910 to 1973.

1. Construct kernel density estimates of the data using the Gaussian and Epanechnikov kernels.
2. Compare the estimates for different choices of bandwidth.
3. Is the estimate more influenced by the type of kernel or the bandwidth?

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 - Newton's method

Consider the density $f(x) = [1 - \cos\{x - \theta\}]/2\pi$ on $0 \leq x \leq 2\pi$, where θ is a parameter between $-\pi$ and π. The following i.i.d. data arise from this density: 3.91, 4.85, 2.28, 4.06, 3.70, 4.04, 5.46, 3.53, 2.28, 1.96, 2.53, 3.88, 2.22, 3.47, 4.82, 2.46, 2.99, 2.54, 0.52, 2.50. We wish to estimate θ.

1. Graph the log-likelihood function between $-\pi$ and π.
2. Find the method of moments estimator of θ.
3. Find the MLE for θ using Newton's method, using the result from 10 as a starting value. What solutions do you find when you start at -2.7 and 2.7?
4. Repeat problem 11 using 200 equally spaced starting values between $-\pi$ and π. The partition the interval into sets of attraction. That is, divide the starting values into separate groups corresponding to the different local modes. Discuss your results.
5. Find two starting values as close together as you can that converge to different solution using Newton's method.

Part II - Gibbs Sampler

Suppose (Y_1, Y_2) are normally distributed with mean $\mu = (0, 0)$ and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

1. Find the full conditional distributions of $Y_1|Y_2$ and $Y_2|Y_1$.
2. Write a Gibbs sampler using the full conditional distributions.
3. Generate 10000 draws from the bivariate normal distribution with $\rho = .7$.
4. Plot estimates of the marginal distributions of Y_1 and Y_2 using the 10000 MCMC draws along with the true distribution. Comment on your findings.
5. Estimate the Effective Sample Size for estimating $E(Y_1)$ and $E(Y_2)$. Comment on your findings.
6. Comment on the mixing properties for your Gibbs sampler. Include at least one plot in support of your comments.

Part III - Flour beetle population

The table below provides counts of a flour beetle population at various points in time. Beetles in all stages of development were counted, and the food supply was careful controlled.

Days (t_i)	0	8	28	41	63	79	97	117	135	154
Beetles ($N^{obs}(t_i)$)	2	47	192	256	768	896	1120	896	1184	1024

An elementary model for population growth is the logistic model given by

$$N(t) = \frac{2K}{2 + (K - 2)\exp\{-rt\}}$$

where $N(t)$ is the population size at time t, r is a growth parameter and K is a parameter that represents the population carrying capacity of the environment. A popular method to estimate the parameters (K, r) is to minimize the objective function

$$\begin{aligned} g(K,r) &= \sum_{i=1}^{n} \left(N(t_i) - N^{obs}(t_i)\right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{2K}{2 + (K-2)\exp\{-rt_i\}} - N^{obs}(t_i)\right)^2 \end{aligned}$$

with respect to K and r. Here n represents the sample size, and t_i take the values $0, 2, 8, 28, \ldots$ (see the table above).

1. Evaluate the function $g(K, r)$ over an appropriately chosen two-dimensional grid. Produce a surface plot using the function `persp()`.
2. Based on the results of part (1), provide estimates $(\hat{K}, \hat{r})$ of the parameters (K, r), which will minimize the function g.
3. In many population modeling applications, an assumption of log-normality is adopted: $\log(N^{obs}(t))$ are independent and normally distributed with mean $\log(N(t))$ and variance $\sigma^2 = 1$. Design a Monte Carlo approach to estimate the sampling distribution of the estimates found in (2). Implement your approach and display histograms for the sampling distributions for $\hat{K}$ and $\hat{r}$.