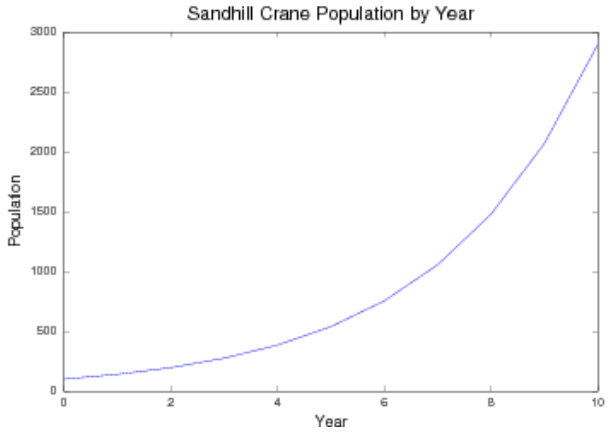
MATH W81: Problem Set 2

Presentations begin Thurs., Jan. 15

All Groups: Let x_n denote the size of the sandhill crane population in a certain region at year n.[1] Let b_n, d_n denote the birth and death rates in year n. Thus, in the absence of any other influences, we expect $x_{n+1} - x_n = b_n x_n - d_n x_n$, or

$$x_{n+1} = (1 + b_n - d_n)x_n \, .$$

(a) As a first pass, assume that the birth and death rates are the same in most years—i.e., that $b_n = 0.5$ and $d_n = 0.1$ holds generally for $n = 0, 1, 2, \ldots$. However, now and then an environmental catastrophe occurs which decreases the birth rate by 40% (i.e., $b_n = (1 - 0.4)(0.5) = 0.3$) and increases the death rate by 25%. These catastrophes occur randomly, but occur, on average, once every 25 years. Write a program in Octave to simulate the progress over 10 years of a sandhill crane population whose initial number is 100. A plot of the resulting population, in the case where no environmental disasters occur, might look like pictured below. Run your program at least twenty times and plot the results together on one graph. Each of your population curves will begin at the point $(0, 100)$. Take note of the distribution of values at year 10. What would the population be if there were no catastrophes? How many different ending values (at year 10) do you see? How many different ending values are possible? Is the ending value a discrete or continuous random variable? Give its pmf/pdf.



We may employ the Unif$(0, 1)$ distribution to simulate the occurrence of catastrophes. Specifically, one might make use of commands like

```
> catastropheOccurs = (rand(1) < 0.04);
> if catastropheOccurs
>   b = 0.3;
>   d = 0.125;
> else
>   b = 0.5;
>   d = 0.1;
> end
```

(b) In the previous model, we assumed birth and death rates changed only occasionally, and then drastically. It would seem more realistic to assume birth and death rates b_n, d_n, $n = 0, 1, \ldots$ were instances of random variables B and D. Let us suppose $B \sim$ Norm$(0.5, 0.03)$ and $D \sim$ Norm$(0.1, 0.08)$. We would then have

$$x_{n+1} = (1 + B - D)x_n \, ,$$

where B and D vary from year to year. Modify your Octave program to account for these new random effects, and produce a time plots (overlaying twenty runs or so) of the population as before. Would it be fair to say that the deterministic model (the one without catastrophes, where $x_{n+1} = 1.4x_n$) describes the *central tendency* of these stochastic models?

(c) You may wish to combine the effects of the previous two parts. I suggest you try to anticipate the results, and then do simulations to see if your expectations are realized.

[1] Douglas Mooney and Randall Swift, *A Course in Mathematical Modeling* (Washington D. C.: Mathematical Association of America, 1999), p. 70 ff.

Group 1: A worker for the Department of Fish and Game is assigned the job of estimating the number of trout in a certain lake of modest size. She proceeds as follows: She catches 100 trout, tags each of them, and puts them back in the lake. One month later, she catches 100 more trout, and notes that 10 of them have tags.

(a) Without doing any fancy calculations, give a rough estimate of the number of trout in the lake.

(b) Let N be the actual number of trout in the lake. Find an expression, in terms of N, for the probability that the worker would catch 10 tagged trout out of the 100 trout that she caught the second time.

(c) Find the value of N which maximizes the expression in part (b). This value is called the *maximum likelihood estimate* for the unknown quantity N.

A census in the United States is an attempt to count everyone in the country. It is inevitable that many people are not counted. The U. S. Census Bureau proposed a way to estimate the correct total number of people in the latest census. Their proposal was as follows: In a given locality, let N denote the actual number of people who live there. Assume that the census counted n_1 people living in this area. Now, another census was taken in the locality, and n_2 people were counted. In addition, n_{12} were counted both times.

(d) Given N, n_1, and n_2, let X denote the number of people counted both times. Find the probability that $X = k$, where k is a fixed positive integer between 0 and n_2.

(e) Now assume that $X = n_{12}$. Find the value of N which maximizes the expression in part (a).[2]

Group 2: (a) While a member of the Mathematics Department at Dartmouth College, Reese Prosser never put money in the (then) 10-cent parking meters of Hanover, NH. He assumed there was a probability of 0.05 he would be caught and ticketed. The cost of the offense: nothing the first time, \$2 the second, and \$5 each subsequent occasion. Under his assumptions, how did the expected cost of parking 100 times without paying the meter compare with the cost of paying each time?[3]

(b) A particular region of South London suffered 537 flying-bomb hits during World War II. The data supplied in the table corresponds to dividing this region of London into 576 smaller regions of 1/4 km^2 each, and recording the number N_k of subregions with exactly k hits.

k	0	1	2	3	4	≥ 5
N_k	229	211	93	35	7	1

Assuming hits in this region are dispersed randomly, find the expected number of subregions (out of the 576) which would have counts of 0, 1, 2, 3, 4, and 5 or more hits.[4]

Group 3: Consider the following behavior. Someone begins at the origin, and arbitrarily chooses to walk one unit to the left or right. She follows this up with another similar choice (without regard to the direction walked at first), and repeats this n times. This is referred to as a *random walk*. We are interested in where she winds up (some integer on the number line). The file `http://www.calvin.edu/~scofield/courses/modeling/octave/randsteps.m` is an Octave function which produces a matrix of specified

[2]Both of these problems taken from Charles M. Grinstead and J. Laurie Snell, *Introduction to Probability: Second Revised Edition* (Providence, RI: American Mathematical Society, 1997), p. 198.

[3]Charles M. Grinstead and J. Laurie Snell, *Introduction to Probability: Second Revised Edition* (Providence, RI: American Mathematical Society, 1997), p. 200.

[4]William Feller, *An Introduction to Probability Theory and Its Applications, Volume I, 3rd Edition* (John Wiley & Sons, Inc., 1968), pp. 160–161.

size whose entries are all ±1, randomly chosen. Use it (or write your own Octave program which does the same thing), along with Octave's `sum()` command, for the following parts.

(a) Run simulations of a random walk with $n = 10$. That is, repeatedly (perhaps 100,000 times or so) find the final destination of a random walk with $n = 10$, storing the results in a vector. Then draw a histogram of relative frequencies for the various destinations $x = -10, -8, \ldots, 10$.

(b) Repeat the previous exercise, but now with $n = 7$.

(c) What probability distributions do the histograms produced in the previous two parts look like? Be as specific as possible, giving not only the name of the distribution family, but also the parameters within that family. Can you justify mathematically the similarity between your random walk data and the corresponding probability distribution?

Group 4: We may never know why the proverbial chicken (or pedestrian) wishes to cross the road. Nevertheless, we might make the following assumptions:

- the flow of automobiles is in a single direction (i.e, a one-way street),
- a certain average amount of time T is required for pedestrians to *walk* across,
- a pedestrian can tell, at least initially, if the time gap between vehicles is sufficiently large ($> T$) to permit a crossing,
- pedestrians who have to wait too long—say, longer than some fixed time τ—may become sufficiently impatient so as to impair judgment about whether there is a sufficient time gap between oncoming cars, resulting in foolish risks.

When the average wait time exceeds τ, it is wise to intervene, installing a traffic signal with pedestrian crossing control.

Assume the number of cars arriving per unit time has a Poisson distribution with rate parameter λ.

(a) Let $G_k, k = 1, 2, \ldots$ denote the inter-arrival time of automobiles starting when a pedestrian arrives, ready to cross a street—i.e., the time between the (k-1)st and kth cars to pass by. Find an expression for $P(G_k > t)$.

Now let X be the random variable whose value is k if $G_k > T$ but each $G_j \le T$ for $j = 1, 2, \ldots, k-1$. Note that X is a discrete random variable. Describe (i.e., give a formula for) its pmf $f_X(k) = P(X = k)$. That is, determine the probabilty that the kth gap will be the first opportunity to cross.

(b) For a discrete random variable X, one calculates the mean using the formula $\mu_X := \sum_k k f_X(k)$, where the index k of summation covers all possible values for the random variable X. Using the fact that

$$\frac{d}{dT}(1 - e^{-\lambda T})^k \;=\; k(1 - e^{-\lambda T})^{k-1}\lambda e^{-\lambda T} \;,$$

find a compact expression for the mean μ_X.

(c) Crossing control is needed when

$$\begin{aligned} \text{Average wait} &= (\text{Avg. gap count just before first crossing possible})(\text{Avg. gap time}) \\ &= (\mu_X - 1)(\text{Avg. gap time while waiting}) \end{aligned}$$

exceeds τ. Observe that the average car inter-arrival time during a pedestrian's wait cannot be larger than T, and hence

$$(\mu_X - 1)T \;\ge\; (\mu_X - 1)(\text{Avg. gap time}) \ge \tau \;.$$

Using your expression for μ_X from the previous part, solve for the threshold rate (i.e., $\lambda \ge$ something) at which crossing control is needed.

(d) Assuming a road is 100 feet across and that pedestrians walk at an average rate of 3.5 ft/second, determine the threshold rate λ (in cars per hour) above which crossing control is needed.[5]

Group 5: The file http://www.calvin.edu/~scofield/data/tab/putting.txt contains results of a study of the putting professional golfers. More specifically, this is a tab-delimited file whose columns give the distance (in feet), the number of putts attempted by professional golfers in the study, and the number of putts made at that distance respectively. Download and save the file in a convenient working directory. Then load it into Octave, and plot the *proportion* of successes against the *distance*. The following commands should achieve this:

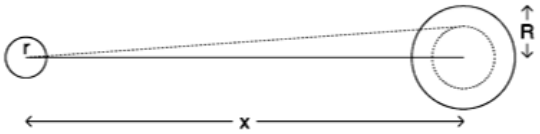
```
> puttdat = load('-ascii', 'putting.txt')
> x = puttdat(:, 1);
> y = puttdat(:, 3) ./ puttdat(:, 2);
> plot(x, y, '*')
```

It is, perhaps, surprising that professional golfers made fewer than 60% of their five-foot putts. Our goal, in this problem, is to find a model (a curve, $y = f(x)$) that fits this data.[6]

(a) There is a distinct upward concavity in the appearance of the plot, so a linear fit ($f(x) = mx+b$) does not seem warranted. Propose some other curve families ($f(x) = ax^p$, or some variant of this family of curves which relegates any vertical asymptote to the left of the y-axis, might be worth a try) which would produce the correct shape. Investigate these curve families, choosing appropriate parameters so as to "best fit" the data. Note that no *simple* curve will pass through all of the data points. The usual way to choose appropriate parameters, when it is analytically possible, is to use optimization techniques from multivariate calculus, requiring a least-squares objective function J(parameters; data) to be minimized. In the case of the curve family $f(x) = f(x; a, p) = ax^p$, this objective function would be

$$J(a, p; \text{data points } (x_k, y_k)) := \sum_{k=1}^{n} (y_k - ax_k^p)^2 .$$

(b) Let us employ the diagram at right as motivation for a different approach to finding a model. The figure shows circles of radius r (the golf ball) and radius R (the hole). Let θ be the angle between the path of the ball and the line segment of length x between the centers of these circles. One might consider θ to be a random variable. A sufficient condition that the ball go in the hole is that the edge of the ball cross the smaller circle of radius $R - r$ pictured inside the hole.



Hypothesize a distribution (or several) for θ. (Like the curve families of part (a), distributions come with parameters as well.) Use the CDF of your distribution to come up with a function $f(x)$, and choose parameters so as to "best fit" the data. Several things you might do to test the validity of your model include

- Plot the model on top of the data points to see if there is fairly good agreement.
- Simulate the selection of θ's from your proposed distribution at various distances (in numbers equal to those attempted at such distances in the data). Look at a plot with this simulated data, and see if it agrees with the model to a similar degree as the real data.

[5] Mooney and Swift, p. 331 ff.

[6] Andrew Gelman and Deborah Nolan, *Teaching Statistics: A Bag of Tricks* (Oxford University Press, 2002).

- Think about what you expect for proportion of successes at the two extreme distances $x = 0$ and $x = \infty$. Does the model yield the correct proportions/probabilities?

(c) Compare/critique the best of your model(s) from part (b) with the best of part (a). Which is a better fit to the data? Is one model more intellectually satisfying than the other? Does either approach rely on assumptions which characterize too simply the actual process of putting in golf? If so, what assumptions are they, and how concerned should we be about them?