Assignment 1 - Experimenting with Statistical Principals*

Turn in a hard copy at the start of class:
Wednesday, February 1

Overview

In this assignment you will experiment with random variation over discrete events.

At some point I did a variation of these experiments by flipping a coin 1000 times are recording the results. Luckily we now have computers, and we scale things up much more easily. Although, you are welcome to use a n-sided die, for appropriate values of n.

As usually, it is highly recommended that you use LaTeX for this assignment. If you do not, you may lose points if your assignment is difficult to read or hard to follow. Find a sample form in this directory: `http://www.cs.utah.edu/~jeffp/teaching/latex/`

1 Q1: Birthday Paradox

Consider a domain of size $n = 1000$.

A: Generate random numbers in the domain $[n]$ until two have the same value. How many random trials did this take? We will use k to represent this value.

B: Repeat the experiment $m = 200$ times, and record for each how many random trials this took. Plot this data as a *cumulative density plot* where the x-axis records the number of trials required k, and the y-axis records the fraction of experiments that succeeded (a collision) after k trials. The plot should show a curve that starts at a y value of 0, and increases as k increases, and eventually reaches a y value of 1.

C: Calculate the empirical expected value of the number of k random trials in order to have a collision. That is, add up all values k, and divide by m.

D: Describe how you implemented this experiment. Would this scale well if instead we have $n = 1000000$ and $m = 10000$? If not, what would you change to make it run faster?

2 Q2: Coupon Collectors

Consider a domain of size $n = 60$.

A: Generate random numbers in the domain $[n]$ until every value $i \in [n]$ has had one random number equal to i. How many random trials did this take? We will use k to represent this value.

*CS 6955 Data Mining; Spring 2012 Instructor: Jeff M. Phillips, University of Utah

B: Make a histogram plot that shows for each i how many times a random number had that value. You should have 60 x values and each should have a height of at least 1.

Report how large was the tallest bar in the chart?

C: Repeat step A for $m = 300$ times, and record for each the value k or how many random trials we required to collect all values $i \in [n]$. Make a cumulative density plot as in 1.B.

D: Calculate the empirical expected value of k.

E: Describe how you implemented this experiment. Would this scale well if instead we have $n = 10000$ and $m = 100000$? If not, what would you change to make it run faster?

3 Q3: Analysis

A: Calculate analytically (using the formulas from class) the number of random trials needed to so there is a collision with probability at least 0.5 when the universe size is $n = 1000$. (Show your work.)

How does this compare to your results from Q1?

B: Calculate analytically (using the formulas from class) the expected number of random trials before all elements are witnessed in a universe of size $n = 60$? (Show your work.)

How does this compare to your results from Q2?

4 BONUS

Consider a domain size n and the coupon collectors problem. Let k represent the number of random trials it takes before you see all n elements.

Show that $\textbf{Pr}[k > 20n \ln n] < 0.1$. (That is, having more than $20n \ln n$ random trials before seeing all n distinct elements happens less than 10 percent of the time.)