Asmt 1: Experimenting with Statistical Principles

Turn in a hard copy at the start of class:
Wednesday, January 30

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 (7 points)

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 and how long it took for $m = 200$ trials.

Estimate how long it would take to run $n = 1000000$ and $m = 10000$, and explain your rationale. (It may be helpful to change n and m and see how the time changes.) If this would not be feasible, how would you change your algorithm to improve the efficiency?

2 Q2: Coupon Collectors (8 points)

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.

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, and how long it took for $n = 60$ and $m = 300$.

Estimate how long it would take to run $n = 10000$ and $m = 10000$, and explain your rationale. If this would not be feasible, how would you change your algorithm to improve the efficiency?

3 Q3: Analysis (5 points)

A: Calculate analytically (using the formulas from class) the number of random trials needed so there is a collision with probability at least 0.5 when the domain 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 domain of size $n = 60$? (Show your work.)

How does this compare to your results from Q2?

4 BONUS (2 points)

Consider a domain size n and let k be the number of random trials run. Let v_i denote the number of trials that have value i. Note that for each $i \in [n]$ we have $\mathbf{E}[v_i] = k/n$. Let $\mu = \max_{i \in [n]} v_i/k$.

Consider some parameter $\gamma \in (0, 1)$. How large does k need to be for $\mathbf{Pr}[|\mu - 1/n| \geq \gamma] \leq 0.1$? That is, how large does k need to be for *all* counts to be within a γ percentage of the average?

How does this change if we want $\mathbf{Pr}[|\mu - 1/n| \geq \gamma] \leq 0.001$?

(Make sure to show your work)