Probability in Computing

CS 237

LECTURE 21

Last time

- Coupon Collector's Problem
- Reservoir Sampling

Today

- Markov inequality
- Chebyshev's inequality

Reminder

- HW 10 is due Thursday

Top Hat question (Join Code: 413437)

For every random variable X, the variance of $3X + 2$ is

A. $3\,\mathrm{Var}(X)$

B. $3\,\mathrm{Var}(X) + 2$

C. $9\,\mathrm{Var}(X)$

D. $9\,\mathrm{Var}(X) + 4$

E. None of the above

Top Hat question (Join Code: 413437)

Pick the first choice that applies:

$$\mathrm{Var}(X + Y + Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z)$$

A. for all random variables X, Y, and Z

B. for all pairwise independent random variables X, Y, and Z

C. for all mutually independent random variables X, Y, and Z

D. for all mutually independent indicator variables X, Y, and Z

E. None of the above.

CS 237

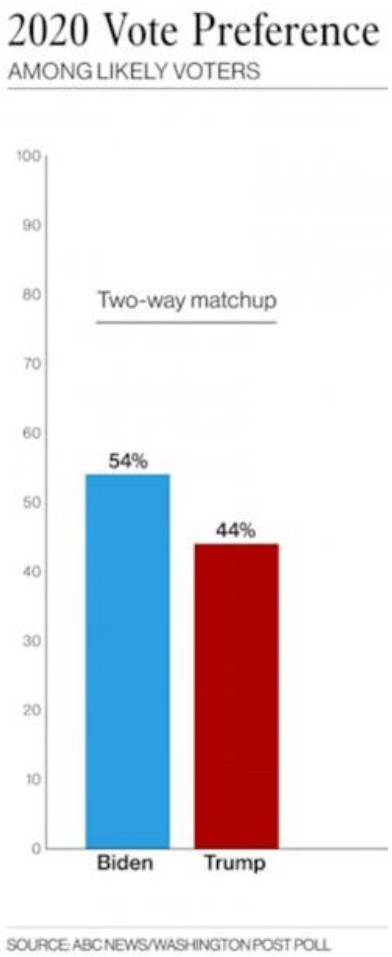
Simulations

- In lecture and in the homework, we have used simulations to empirically estimate probabilities, expectations, and distributions
- In each of those cases, we used many repeated trials (typically 10,000 trials)
- Our goal this week is to understand questions like:
  - How accurate were our estimates?
  - How do we decide how many trials/samples we need?
  - Were 10,000 trials too few, too many, or just right?

Estimation by Sampling

We will use polling as a running example

- Suppose we have an upcoming election with two candidates: candidate A and candidate B
- Let p be the fraction of the voters that support candidate A
- The job of a pollster is to estimate this unknown fraction p
- How can the pollster proceed?



Estimation by Sampling

- Approach 1: Pollster calls every voter and asks them which candidate they support
  - **Pro:** the pollster will get a perfect estimate
  - **Con:** the pollster will need to call hundreds of millions of people, this is nearly impossible and could take years
- Approach 2: Pollster calls a small sample of voters and asks them which candidate they support
  - **Pro:** the number of people is much smaller
  - **Con:** the estimate could be very inaccurate

Estimation by Sampling: Polling

The pollster uses the following polling algorithm:

1. Choose a sample size n
2. Sample n people independently and uniformly at random with replacement from the entire population
3. For each sampled person, ask them which candidate they support (we assume they answer truthfully)
4. Use the fraction of people in the sample that support candidate A as the estimate for the true fraction

Estimation by Sampling: RVs

- For each $i \in \{1,2,\ldots,n\}$, define:

$$X_i = \begin{cases} 1 \text{ if } i^{\text{th}} \text{ person in the sample supports A} \\ 0 \quad \text{otherwise} \end{cases}$$

- The pollster's estimate is:

$$P = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Our goal is to understand how ``close'' the estimate P is to the actual fraction of voters that support candidate A

Expectation of Estimate P

- Fraction of the population supporting candidate A is $\boldsymbol{p}$
- The pollster's estimate is: $\boldsymbol{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

  where $X_i = \begin{cases} 1 \text{ if } i^{\text{th}} \text{ person in the sample supports A} \\ 0 \quad \text{otherwise} \end{cases}$
- What are the distribution, expectation and variance of X_i?

  $X_i \sim$ $\quad\quad$ $\mathbb{E}(X_i) =$ $\quad\quad$ $\text{Var}(X_i) =$
- By linearity of expectation,

$$\mathbb{E}(P) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) =$$

$\mathbb{E}(P)$ is exactly the unknown fraction p we wanted to estimate

Variance of Estimate P

- Fraction of the population supporting candidate A is $\boldsymbol{p}$
- The pollster's estimate is: $\boldsymbol{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$,

  where $X_i = \begin{cases} 1 \text{ if } i^{\text{th}} \text{ person in the sample supports A} \\ 0 \quad \text{otherwise} \end{cases}$

  $X_i \sim$ $\quad\quad\quad\quad\quad\quad$ $\mathbb{E}(X_i) =$ $\quad\quad\quad\quad\quad\quad$ $\text{Var}(X_i) =$
- Since X_i are independent,

$$\text{Var(P)} = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) =$$

The larger n is, the smaller Var(P) is.

CS 237

Estimation by Sampling

- The pollster's estimate is correct in expectation
- Next we introduce the tools we will need to understand how much it can deviate from its expectation

Markov Inequality

Theorem (Markov Inequality)

Let X be a random variable taking only nonnegative values. Then, for all $a > 0$,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Useful when $a > \mathbb{E}(X)$

Andrei Markov
[1856-1922]

Proof: Let $a > 0$.

image source https://en.wikipedia.org/wiki/Andrey_Markov

Markov Inequality: Proof

Markov Inequality: Corollary

Theorem (Markov Inequality)

Let X be a random variable taking only nonnegative values. Then, for all $a > 0$,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Andrei Markov
[1856-1922]

Corollary (Variant of Markov Inequality)

Let X be a random variable taking only nonnegative values. Then, for all $b > 1$,

$$\Pr(X \geq b \cdot \mathbb{E}(X)) \leq \frac{1}{b}.$$

Markov Inequality: Assumption

Theorem (Markov Inequality)

Let X be a RV taking only nonnegative values. Then, for all $a > 0$,

$$\Pr(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Andrei Markov
[1856-1922]

- The nonnegativity assumption is necessary.
- Consider, for example,

$$X = \begin{cases} 1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases}$$

Then $\Pr(X \geq 1) = \frac{1}{2} \not\leq \mathbb{E}(X) = 0$

Markov Inequality: Polling

Corollary (Markov Inequality)

Let X be a RV taking only nonnegative values. Then, for all $b > 1$,

$$\Pr(X \geq b \cdot \mathbb{E}(X)) \leq \frac{1}{b}.$$

Andrei Markov
[1856-1922]

- By Markov inequality, applied to the polling estimate,

$$\Pr(P \geq 1.1p) \leq$$
$$\Pr(P \geq 2\,p) \leq$$
$$\Pr(P \geq b\,p) \leq$$

Markov Inequality: Polling

- The guarantees we obtained using Markov inequality seem rather weak
- The guarantee is the same regardless of whether we poll 1 person or 225,000,000 people
- Can we do better?

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

Let X be a random variable. For all $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Pafnuty Chebyshev

[1821-1894]

image source https://www.britannica.com/biography/Pafnuty-Lvovich-Chebyshev

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

Let X be a random variable. For all $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Proof: $\Pr(|X - \mathbb{E}[X]| \geq a) = \Pr(\underbrace{(X - \mathbb{E}[X])^2}_{Y \geq 0} \geq a^2)$

$$\leq \frac{\mathbb{E}(Y)}{a^2} \qquad \text{(by Markov)}$$

$$= \frac{\mathbb{E}((X - \mathbb{E}[X])^2)}{a^2}$$

$$= \frac{\mathrm{Var}(X)}{a^2}$$

Chebyshev's Inequality: Polling

Theorem (Chebyshev's Inequality)

Let X be a random variable. For all $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Pafnuty Chebyshev

[1821-1894]

- By applying Chebyshev's inequality to the polling estimate,

$\Pr(|P - p| \geq a) \leq$

Chebyshev's Inequality: Polling

- The Chebyshev inequality bound allows us to determine how many people to poll
- Suppose we want the estimate to be within 0.04 of p with probability at least 0.95

Want:

$$\Pr(|P - p| \leq 0.04) \geq 0.95$$
$$\Leftrightarrow \Pr(|P - p| > 0.04) \leq 0.05$$

Chebyshev:

$$\Pr(|P - p| > 0.04) \leq \frac{1}{4 \cdot (0.04)^2 \cdot n}$$

Thus it suffices to poll $n = 3125$ voters

CS 237

Estimation by Sampling

- The approach we studied in the context of polling can be used in a wide range of settings
- A common scenario in CS and beyond is that we want to estimate the expectation of a distribution using sampling
- Our earlier analysis works equally well for this setting: we take several samples from the distribution and we use their average as our estimate for the expectation
- Chebyshev's inequality then tells us how close our estimate is to the actual expectation, and it also tells us how many samples we need