CANDIDATE POPULARITY

Favorable  | Unfavorable

Donald Trump: 37% Favorable, 60% Unfavorable
Hillary Clinton: 44% Favorable, 53% Unfavorable
Bernie Sanders: 51% Favorable, 38% Unfavorable

Source: ABC News/Washington Post poll
At many times 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 today (and next lecture) 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, let us call them A and 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?
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
ESTIMATION BY SAMPLING

- Approach 1: Pollster calls every voter and asks them which candidate they support
- 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
The pollster uses the following polling algorithm:

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

› For each $i \in \{1, 2, \cdots, n\}$, define:

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

› The pollster’s estimate is a random variable:

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

› Our goal is to understand how “close” the estimate $P$ is to the actual fraction of voters that support candidate A
What are the distribution, expectation and variance of $X_i$?

$X_i \sim$ \quad $\mathbb{E}(X_i) =$ \quad $Var(X_i) =$

By linearity of expectation:

$\mathbb{E}(P) = \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) =$
Since $X_i$ are mutually independent:

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

- Developing the intuition: What fraction of students can get at least twice the average on an exam? at least thrice the average?
Theorem (Markov Inequality):

If $X$ is a non-negative random variable, then for all $a > 0$

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

- Theorem (Markov Inequality):

  If $X$ is a non-negative random variable, then for all $a > 0$

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

- Corollary:

  If $X$ is a non-negative random variable, then for all $c \geq 1$

  $$ Pr(X \geq c \cdot E(X)) \leq \frac{1}{c} $$
Markov Inequality

Theorem (Markov Inequality):
If $X$ is a non-negative random variable, then for all $a > 0$

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

Note that the non-negativity assumption is necessary:

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

$$Pr(X \geq 1) = \frac{1}{2} \neq \mathbb{E}(X) = 0$$
MARKOV INEQUALITY

- Markov Inequality:

If $X$ is a non-negative random variable, then for all $c \geq 1$

\[ Pr(X \geq c \cdot \mathbb{E}(X)) \leq \frac{1}{c} \]

- By applying it to the polling estimate:
BU Terriers win each game independently with probability $\frac{2}{3}$, otherwise, they lose the game. Let $X$ be the number of their losses in $n$ games.

- What bound on $\Pr(X \geq \frac{n}{2})$ can you get using Markov inequality?

A. $\leq \frac{1}{3}$
B. $\leq \frac{1}{2}$
C. $\geq \frac{1}{2}$
D. $\leq \frac{2}{3}$
E. None of the above
MARKOV INEQUALITY

- The guarantees that we obtained using Markov’s inequality seem rather weak
- We obtain the same guarantee regardless of whether we poll 1 person or 225,000,000 people
- Can we do better?
**Chebyshev Inequality**

- **Theorem (Chebyshev Inequality):** Pafnuty Chebyshev

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

$$Pr(|X - \mathbb{E}(X)| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$
BU Terriers win each game independently with probability 2/3. Let $X$ be the number of their losses in $n$ games. What bound on $\Pr(X \geq \frac{n}{2})$ can you get using Chebyshev’s inequality?

A. $\leq \frac{1}{3}$

B. $\leq \frac{\text{const}}{n}$

C. $\geq \frac{\text{const}}{n^2}$

D. $\leq \frac{\text{const}}{n^3}$

E. None of the above
CHEBYSHEV INEQUALITY

- Theorem (Chebyshev Inequality): Pafnuty Chebyshev

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

$$Pr(|X - \text{Ex}(X)| \geq a) \leq \frac{\text{Var}(X)}{a^2}$$

- By applying it to the polling estimate:
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:

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

<table>
<thead>
<tr>
<th>Want:</th>
<th>Chebyshev:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Pr(</td>
<td>P - p</td>
</tr>
<tr>
<td>$Pr(</td>
<td>P - p</td>
</tr>
</tbody>
</table>
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:
\[
\Pr(|P - p| \leq 0.04) \geq 0.95 \\
\Pr(|P - p| > 0.04) \leq 0.05
\]

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

Thus it suffices to poll $n = 3125$ voters.
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.