1. **Playing Pollster**

As an expert in probability, the staff members at the Daily Californian have recruited you to help them conduct a poll to determine the percentage p of Berkeley undergraduates that plan to participate in the student sit-in. They've specified that they want your estimate $\hat{p}$ to have an error of at most ε with confidence $1-\delta$. That is,

$$P(|\hat{p}-p| \leq \varepsilon) \geq 1-\delta.$$

Assume that you've been given the bound

$$P(|\hat{p}-p| \geq \varepsilon) \leq \frac{1}{4n\varepsilon^2},$$

where n is the number of students in your poll.

(a) Using the formula above, what is the smallest number of students n that you need to poll so that your poll has an error of at most ε with confidence $1-\delta$?

Solution: We know we need to have

$$P(|\hat{p}-p| \leq \varepsilon) \geq 1-\delta.$$

Subtracting both sides from 1, it follows that we must have

$$P(|\hat{p}-p| > \varepsilon) \leq \delta.$$

Therefore if we choose n such that

$$\frac{1}{4n\varepsilon^2} \leq \delta,$$

we will have

$$P(|\hat{p}-p| \geq \varepsilon) \leq \delta,$$

and since $P(|\hat{p}-p| > \varepsilon) \leq P(|\hat{p}-p| \geq \varepsilon)$, this will meet the requirement that

$$P(|\hat{p}-p| > \varepsilon) \leq \delta.$$

Thus we must have that

$$\begin{aligned}\frac{1}{4n\varepsilon^2} &\leq \delta \\ \frac{1}{n} &\leq 4\varepsilon^2\delta \\ n &\geq \frac{1}{4\varepsilon^2\delta}.\end{aligned}$$

(b) At Berkeley, there are about 26,000 undergraduates and about $10,000$ graduate students. Suppose you only want to understand the frequency of sitting-in for the undergraduates. If you want to obtain an estimate with error of at most 5% with 98% confidence, how many undergraduate students would you need to poll? Does your answer change if you instead only want to understand the frequency of sitting-in for the graduate students?

Solution: Plugging in to the bound you found above, you get that $n \geq 5000$. The answer is the same for graduate students; the size of the population does not affect the number of samples you need.

(c) It turns out you just don't have as much time for extracurricular activities as you thought you would this semester. The writers at the Daily Californian insist that your poll results are reported with at least 95% confidence, but you only have enough time to poll 500 students. Based on the bound above, what is the worst-case error with which you can report your results?

Solution: If you only have time to poll 500 people and want to report your results with 95% confidence, you must report that the error in your estimate is at most 10%. You can find this by plugging in $\frac{1}{4 \cdot 500 \cdot \varepsilon^2} = .05$ and solving for ε.

2. **Vegas**

On the planet Vegas, everyone carries a coin. Many people are honest and carry a fair coin (heads on one side and tails on the other), but a fraction p of them cheat and carry a trick coin with heads on both sides. You want to estimate p with the following experiment: you pick a random sample of n people and ask each one to flip his or her coin. Assume that each person is independently likely to carry a fair or a trick coin.

(a) Given the results of your experiment, how should you estimate p?

(b) How many people do you need to ask to be 95% sure that your answer is off by at most 0.05?

Solution:

(a) You count the fraction of heads that you see. If q is this fraction then $q \simeq 1/2(1-p) + p = 1/2 + p/2$. So you declare p to be $2q-1$.

(b) We want $2q-1$ to be within 0.05 of its mean. This means that q should be within 0.025 of its mean, or the sum should be within $0.025n$ of its mean. The variance of each coin flip is $q(1-q)$, therefore Chebyshev tells us that

$$\Pr[|\sum_{i=1}^{n} X_i - qn| \geq 0.025n] \leq q(1-q)/(0.025n)^2$$

We have $q(1-q) < 1/4$. So for the probability to be bounded by 5% we can have $n = 90 > \sqrt{8000}$.

3. **Working with the Law of Large Numbers**

(a) A fair coin is tossed and you win a prize if there are more than 60% heads. Which is better: 10 tosses or 100 tosses? Explain.

Solution: 10 tosses.

(b) A fair coin is tossed and you win a prize if there are more than 40% heads. Which is better: 10 tosses or 100 tosses? Explain.

Solution: 100 tosses. Based on the first part, consider the inverse of the event "more than 60% heads" and the symmetry of heads and tails.

(c) A coin is tossed and you win a prize if there are between 40% and 60% heads. Which is better: 10 tosses or 100 tosses? Explain.

Solution: 100 tosses. Based on the first part, consider the union of the events "more than 60% heads" and "more than 60% tails" ("less than 40% heads").

(d) A coin is tossed and you win a prize if there are exactly 50% heads. Which is better: 10 tosses or 100 tosses? Explain.

Solution: 10 tosses. Compare the probability of getting equal number of heads and tails between $2n$ and $2n+2$ tosses.

$$\begin{aligned}
\Pr[n \text{ heads in } 2n \text{ tosses}] &= \binom{2n}{n}/2^{2n} \\
\Pr[n+1 \text{ heads in } 2n+2 \text{ tosses}] &= \binom{2n+2}{n+1}/2^{2n+2} \\
&= \frac{(2n+2)!}{(n+1)!(n+1)!} \cdot \frac{1}{2^{2n+2}} \\
&= \frac{(2n+2)(2n+1)2n!}{(n+1)(n+1)n!n!} \cdot \frac{1}{2^{2n+2}} \\
&= \frac{2n+2}{n+1} \cdot \frac{2n+1}{n+1} \binom{2n}{n} \cdot \frac{1}{2^{2n+2}} \\
&< \left(\frac{2n+2}{n+1}\right)^2 \binom{2n}{n} \cdot \frac{1}{2^{2n+2}} \\
&= 4\binom{2n}{n} \cdot \frac{1}{2^{2n+2}} = \binom{2n}{n}/2^{2n} = \Pr[n \text{ heads in } 2n \text{ tosses}]
\end{aligned}$$

The larger n is, the less probability we'll get 50% heads. $\square$