Question 3

(a)

Compute a value of m so that the result of the poll is incorrect with probability at most 1%? Use the Hoeffding/Chernoff bounds and show your work.

Answer

We use Theorem 16.2 to determine the required number of samples, n, to achieve the desired confidence in the polling result.
Theorem 16.2 states: Let $W \in \{0, 1\}$, and let $X_1, \ldots, X_n \in \{0, 1\}$ be independent random variables such that $\Pr[X_i = W] \geq \frac{1}{2} + \varepsilon$. Then:

$$\Pr[\text{Majority}(X_1, \ldots, X_n) = W] \geq 1 - 2e^{-2\varepsilon^2 n}$$

Given that we want the probability that the poll result is incorrect to be at most 1%, i.e.,

$$\Pr[\text{Majority}(X_1, \ldots, X_n) \neq W] \leq 0.01,$$

by Theorem 16.2, we require:

$$1 - 2e^{-2\varepsilon^2 n} \geq 0.99.$$

Rearranging, we find:

$$2e^{-2\varepsilon^2 n} \leq 0.01.$$

Taking natural logarithms on both sides, we obtain:

$$\begin{aligned} \ln(2) + \ln(e^{-2\varepsilon^2 n}) &\leq \ln(0.01), \\ \ln(2) - 2\varepsilon^2 n &\leq \ln(0.01). \end{aligned}$$

Solving for n, we get:

$$\begin{aligned} -2\varepsilon^2 n &\leq \ln(0.01) - \ln(2), \\ n &\geq \frac{\ln(0.01) - \ln(2)}{-2\varepsilon^2}. \end{aligned}$$

Hence, the required sample size n can be calculated by substituting a specific value for ε. For example, assuming $\varepsilon = 0.05$, we have:

$$n \geq \frac{\ln(0.01) - \ln(2)}{-2 \times 0.05^2}.$$

The precise calculation yields $n \approx 1059.663$. Therefore, to ensure that the poll's result is incorrect with a probability of at most 1%, the number of samples m **needs to be greater than** 1060**.**

(b)

Let n be the number of people in the population, ε be defined such that $\left(\frac{1}{2}+\varepsilon\right)\cdot n$ prefer A to B, and let δ be the desired accuracy (so the probability the result is incorrect is at most δ). Write your bound m as a function of n, ε, and δ.

- If the number of people in the population increased by a factor of 10, how would that affect m?
- If ε decrease by a factor of 2, how would that affect m?
- If we want to increase our confidence by a factor of 10 ($\delta' = \delta/10$), how would that change m?
- If $\varepsilon = 1/n$ (so 1 person would be the deciding vote), what would this imply about m given your bound from above?

Answer

Derivation of Minimum Sample Size m

Given:

- n is the total number of people in the population.
- ε such that $\left(\frac{1}{2}+\varepsilon\right)\cdot n$ people prefer A over B.
- δ is the desired accuracy, such that the probability that the result is incorrect is at most δ.

Applying to Polling

In polling, the X_i are Bernoulli trials where $X_i = 1$ if the i-th respondent prefers A over B, and $\mu = \frac{1}{2}+\varepsilon$ represents the proportion of the population that prefers A over B. The inequality becomes from **Thoerem 12.6**:

$$\Pr\left[\text{Majority incorrectly predicted}\right] \leq 2\exp(-2m\varepsilon^2)$$

To meet the requirement that the probability of an incorrect prediction is at most δ, we set:

$$2\exp(-2m\varepsilon^2) \leq \delta$$

Solving for m, we get:

$$\exp(-2m\varepsilon^2) \leq \frac{\delta}{2},$$

$$-2m\varepsilon^2 \leq \ln\left(\frac{\delta}{2}\right),$$

$$m \geq \frac{\ln\left(\frac{2}{\delta}\right)}{2\varepsilon^2}.$$

Conclusion

Thus, the minimum sample size m necessary to ensure that the polling result is incorrect with a probability of at most δ is given by:

$$m \geq \frac{\ln\left(\frac{2}{\delta}\right)}{2\varepsilon^2}.$$

Effects of Changes in Parameters on the Minimum Required Sample Size m

Given the formula for the minimum required sample size:

$$m \geq \frac{\ln\left(\frac{2}{\delta}\right)}{2\varepsilon^2}$$

Effect of Increasing the Population Size by a Factor of 10

The formula for m does not directly depend on the total population size n. Thus, increasing n by any factor does not affect m, as m is solely a function of ε and δ.

Effect of Decreasing ε by a Factor of 2

Decreasing ε impacts m significantly. If ε is halved ($\varepsilon' = \varepsilon/2$):

$$m' \geq \frac{\ln\left(\frac{2}{\delta}\right)}{2(\varepsilon/2)^2} = 4 \times \frac{\ln\left(\frac{2}{\delta}\right)}{2\varepsilon^2}$$

This implies that m increases by a factor of 4, illustrating the inverse square relationship between ε and m.

Effect of Increasing Confidence by a Factor of 10 ($\delta' = \delta/10$)

To achieve a tenfold increase in confidence ($\delta' = \delta/10$), we modify m:

$$m' \geq \frac{\ln\left(\frac{2}{\delta/10}\right)}{2\varepsilon^2} = \frac{\ln(20/\delta)}{2\varepsilon^2}$$

Considering $\ln(20/\delta) = \ln(2/\delta) + \ln(10)$, the required m increases due to the added logarithmic term $\ln(10) \approx 2.302$.

Setting $\varepsilon = 1/n$

When $\varepsilon = 1/n$, implying one person can swing the preference:

$$m \geq \frac{\ln\left(\frac{2}{\delta}\right)}{2(1/n)^2} = \frac{n^2 \ln\left(\frac{2}{\delta}\right)}{2}$$

Here, m increases quadratically with n, suggesting that for large populations, the required sample size becomes impractically large, reflecting the sensitivity of m to small changes in ε.

(c)

In practice, what might be wrong with the above assumptions (i.e. why might we not we use polls to run our elections)?

Answer

Uniform Independent Distribution: The formula assumes that the preferences of the voters (represented by the random variables) are independent and **identically distributed.** In reality, voters' decisions can be correlated due to **shared information, social influences, demographic factors** such like Locations, neighborhoods and cities that correspond to culutre and political views, and other regional variables.

All of these factors can make the Distribution not a **Uniform Independent Distribution**.