A **Polling** We want to estimate the percentage of Republicans in the country by conducting a poll. What is the minimum number of people we must sample to be confident of getting an estimate with error at most 5% with a probability of greater than 99% ? (Assume that every person in the population is chosen uniformly and randomly.)

1. First, compute the minimum sample size under the assumption that exactly half the population is Republican.

2. Let S denote the percent of the sample who are Republican and let P denote the percent of the population who are Republican. Given that, for all ϵ,

$$Pr[|S - P| > \epsilon]$$

is maximized when $P = 1/2$ what is the minimum required sample size?

B **Catching fish** There is a lake with 1 million fish in it. We conduct an experiment on the fish in two phases. In phase 1 we catch A fish, tag them and release them back into the lake. In phase 2 we repeatedly catch and release the fish - we do this B times. What is the minimum total number $(A + B)$ of fish we should catch to make sure that we catch a tagged fish (during phase 2) with probability greater than 99%. Assume that every fish in the lake is caught uniformly at random.

Solutions

A **Polling** 1. Let us assume that the correct estimate of the percentage of Republicans is 50%. Let our sample size be n. Let X be the random variable that denotes the number of people in our sample who are Republicans. Then we want to find the minimum n such that

$$Pr[|\frac{X}{n} - .5| > 0.05] < 0.01$$

We have that

$$\begin{aligned} Pr[|\frac{X}{n} - .5| > 0.05] &= Pr[X > 0.55n] + Pr[X < 0.45n] \\ &= 2 * Pr[X < 0.45n] && \text{by symmetry, since } p = 1/2. \\ &\leq 2 * (\tfrac{1-0.45}{1-2*0.45}) 2^{-n} \binom{n}{0.45n} && \text{geometric series} \\ &\leq 11 * 2^{-n} 2^{H(0.45)n} && \text{Stirling} \\ &= 11 * 2^{-(1-H(0.45))n} \end{aligned}$$

Setting the above to be less than 0.01 we find that

$$11 * 2^{-(1-H(0.45))n} < 0.01 \Rightarrow n > \frac{\log_2 1100}{1 - H(0.45)}.$$

i.e. $n > 1398$.

2. Now, note that, for all ϵ, $Pr[|\frac{X}{n} - \mu| > \epsilon]$ is maximized when $\mu = \frac{1}{2}$. (Take this fact for granted.)

From this we conclude that a sample size of 1399 suffices even when we drop the assumption that Republicans form exactly 50% of the population.

B **Catching fish** The probability we don't catch even one tagged fish during phase 2 is $(1 - \frac{A}{10^6})^B$. We want this probability to be less than 0.01. Now,

$$(1 - \frac{A}{10^6})^B < e^{-(\frac{AB}{10^6})}$$

. Hence if we assure that

$$e^{-(\frac{AB}{10^6})} < 0.01$$

we are done. This implies that $AB > 4.6 * 10^6$. $A + B$ is minimized when $A = B$ for a given AB. Hence the minimum total number of fish we should catch is $2 * \sqrt{4.6 * 10^6} \sim 4292$