Lecture 36

Andrei Antonenko

April 27, 2005

1 Polling revisited

Let us revisit the polling problem. We poll n voters and record the fraction M_n of those polled who are in favor of a particular candidate. If p is the fraction of the entire voter population that supports this candidate, then

$$M_n = \frac{X_1 + \cdots + X_n}{n}, \tag{1}$$

where the X_i are independent Bernoulli random variables with parameter p. In particular, M_n has mean p and variance $p(1-p)/n$. By the normal approximation, $X_1 + \cdots + X_n$ is approximately normal, and therefore M_n is also approximately normal.

We are interested in the probability $P(|M_n - p| \geq \epsilon)$ that the polling error is larger than some desired accuracy ϵ. Because of the symmetry of the normal PDF around the mean, we have

$$P(|M_n - p| \geq \epsilon) \approx 2P(M_n - p \geq \epsilon).$$

The variance $p(1-p)/n$ of $M_n - p$ depends on p and is therefore unknown. We note that the probability of a large deviation from the mean increases with the variance. Thus, we can obtain an upper bound on $P(Mn - p \geq \epsilon)$ by assuming that $M_n - p$ has the largest possible variance, namely, $1/4n$. To calculate this upper bound, we evaluate the standardized value

$$z = \frac{\epsilon}{1/(2\sqrt{n})},$$

and use the normal approximation

$$P(M_n - p \geq \epsilon) \leq 1 - \Phi(z) = 1 - \Phi(2\epsilon\sqrt{n}).$$

For instance, consider the case where $n = 100$ and $\epsilon = 0.1$. Assuming the worst-case variance, we obtain

$$\begin{aligned} P(|M_{100} - p| \geq 0.1) &\approx 2P(M_n - p \geq 0.1) \\ &\approx 2 - 2\Phi(2 \cdot 0.1 \cdot \sqrt{100}) = 2 - 2\Phi(2) = 2 - 2 \cdot 0.977 = 0.046. \end{aligned}$$

This is much smaller (more accurate) than the estimate that was obtained using the Chebyshev inequality.

We now consider a reverse problem. How large a sample size n is needed if we wish our estimate M_n to be within 0.01 of p with probability at least 0.95?

Assuming again the worst possible variance, we are led to the condition

$$2 - 2\Phi(2 \cdot 0.01 \cdot \sqrt{n}) \le 0.05,$$

or

$$\Phi(2 \cdot 0.01 \cdot \sqrt{n}) \ge 0.975.$$

From the normal tables, we see that $\Phi(1.96) = 0.975$, which leads to

$$2 \cdot 0.01 \cdot \sqrt{n} \ge 1.96,$$

or

$$n \ge \frac{(1.96)^2}{4 \cdot (0.01)^2} = 9604.$$

This is significantly better than the sample size of 50,000 that we found using Chebyshev's inequality.