The Pollster's Problem

Anirudh Srinivasan

IIT Hyderabad

July 2, 2021

Problem Definition

The Pollster's Problem

The Boss asks pollster to calculate the number of people to be randomly picked as a sample of population to estimate the value of the fraction of population that will vote "yes" in a referendum (denoted by 'p')

Pollster's Initial Approach and its Limitations

Steps

1. 'n' people are picked randomly, uniformly and independently over the population and their answers are recorded in indicator random variables

$$\Pr(X_i) = \begin{cases} 1, & \text{if yes} \\ 0, & \text{if no} \end{cases} \tag{1}$$

2. Fraction of "yes" in our sample = $M_n = \frac{X_1+X_2+\cdots+X_n}{n}$
   This would be a reasonable estimate for p, however, there is no way to find the exact value of p on the basis of finite and random poll. Hence, there is going to be an error in the estimation of p.
3. Also, there is no way of guaranteeing the estimate of p with a small error and with certainity as the people polled might not be representative of the true population.

Revised Problem Definition

The Pollster's Problem

- The Boss asks Pollster to calculate the number of people to be randomly picked as a sample of population to estimate the value of the fraction of population that will vote "yes" in a referendum (denoted by 'p') with minimal probability of low accuracy.
- The desired specifications have 2 parameters:
  1. Accuracy of 99% [1% margin of error] and
  2. Probability of an error greater than the margin of error is less than 5% [95% confident with the accuracy that is going to be achieved]

Specifications

$$\Pr\left(|M_n - p| \geq 0.01\right) \leq 0.05 \tag{2}$$

Theorem

The Chebyshev's Inequality

1. For a random variable X with finite mean μ and variance σ^2,

$$\Pr\left(|X - \mu| \geq c\right) \leq \frac{\sigma^2}{c^2} \tag{3}$$

   where c is any positive real number.

2. Intuitively, this states that X is unlikely to be too far from the mean if the variance is small.

Solution

Consider the i.i.d random variables X_is that act as an indicator that answer is "yes" in the poll. These follow Bernoulli's Distribution with:

$$\Pr(X_i = x) = \begin{cases} p, & \text{for x} = 1 \\ 1-p, & \text{for x} = 0 \end{cases} \tag{4}$$

$$\mu = E[X_i] = \sum_{\forall x} \Pr(X_i = x) \times x \tag{5}$$

$$\mu = p \times 1 + (1-p) \times 0 = p \tag{6}$$

As X_i is bernoulli random variable, $E[X_i] = E[(X_i)^2] = p$ as x={0,1} is the solution set of $x^2 = x$.

$$\sigma^2 = E[(X_i)^2] - (E[X_i])^2 \tag{7}$$

$$= p - p^2 = p(1-p) \tag{8}$$

Solution

Let M_n be the sample mean of the i.i.d random variables X_is.

$$M_n = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} \tag{9}$$

$$E[M_n] = \frac{E[X_1 + X_2 + X_3 + \cdots + X_n]}{n} = \frac{n\mu}{n} = \mu = p \tag{10}$$

$$var(M_n) = \frac{var(X_1 + X_2 + X_3 + \cdots + X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} \tag{11}$$

Solution - Method 1

By applying Chebyshev's Inequality for M_n, we get:

$$\Pr\left(|M_n - \mu| \geq \epsilon\right) \leq \frac{var(M_n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \leq \frac{1}{4n\epsilon^2} \tag{12}$$

Comparing this with the desired specifications, we get:

$$\epsilon = 0.01 \text{ and } \frac{1}{4n\epsilon^2} = 0.05 \tag{13}$$

Solving this, we get: n = 50000

Theorem

The Central Limit Theorem

For a i.i.d random variables $X_1, X_2, \ldots, X_n$ with finite mean μ and variance σ^2, the random variable given by:

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} \text{ (where: } S_n = X_1 + X_2 + \cdots + X_n) \tag{14}$$

follows approximately standard normal distribution for larger n.

Solution - Method 2

We have to rewrite the event $|M_n - p| \geq 0.01$ with an equivalent way involving the standardized normal variable Z_n so that we can use Central Limit Theorem to approximate $\Pr\left(|M_n - p| \geq 0.01\right)$

$$|M_n - p| \geq 0.01 \implies \left|\frac{S_n}{n} - p\right| \geq 0.01 \tag{15}$$

$$\implies \left|\frac{S_n - np}{n}\right| \geq 0.01 \implies \left|\frac{S_n - np}{\sqrt{n}}\right| \geq 0.01\sqrt{n} \tag{16}$$

$$\implies \left|\frac{S_n - np}{\sqrt{n}\sigma}\right| \geq \frac{0.01\sqrt{n}}{\sigma} \implies |Z_n| \geq \frac{0.01\sqrt{n}}{\sigma} \tag{17}$$

$$\implies |Z_n| \geq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}} \tag{18}$$

Solution - Method 2

$$\Pr\left(|M_n - p| \geq 0.01\right) \approx \Pr\left(|Z| \geq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) \tag{19}$$

$$\Pr\left(|Z| \geq \frac{0.01\sqrt{n}}{\sqrt{p(1-p)}}\right) \leq \Pr\left(|Z| \geq 0.02\sqrt{n}\right) \tag{20}$$

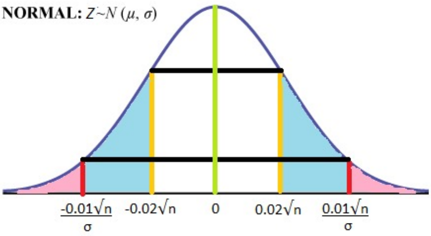

Figure: Normal Distribution of Z

Solution - Method 2

$$\Pr\left(|Z| \geq 0.02\sqrt{n}\right) = 2 \times \Pr\left(Z \geq 0.02\sqrt{n}\right) \tag{21}$$

$$\Pr\left(Z \geq 0.02\sqrt{n}\right) = 1 - \phi(0.02\sqrt{n}) \tag{22}$$

$$\implies \Pr\left(|M_n - p| \geq 0.01\right) \leq 2(1 - \phi(0.02\sqrt{n})) \tag{23}$$

Comparing this with the desired specifications, we get:

$$2(1 - \phi(0.02\sqrt{n})) = 0.05 \tag{24}$$

$$\implies \phi(0.02\sqrt{n}) = 0.975 \tag{25}$$

$$\implies 0.02\sqrt{n} = 1.96 \tag{26}$$

Solving this, we get: n = 9604

Solution - Method 2

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Figure: Z Table - Standard Normal Distribution Table