0.1 Determine the valid sample size

People trend to believe the results of surveys, but are they true? The sample size may be a key to answer the question.

Ground Truth

Let's start with a simple model of 'How do people support a candidate?'. The supporting rate is a random variable that fits

$$R_{support} \sim \mathcal{N}(\mu, \sigma^2)$$

where μ and σ^2 are unavailable.

Population Survey The supporting rate is so important that they want to know it using *LARGE* surveys. It is like asking every body. The number of supporters follows binominal distribution

$$p(n) = (N, n) \cdot r^n \cdot (1 - r)^(N - n)$$

where r refers the ground truth of supporting rate, N is the population and n is the number of supporters.

Then, the sample expectation and sample variance of n is direct

$$E(n) = N \cdot r \tag{1}$$
$$D(n) = N \cdot r \cdot (1 - r) \tag{2}$$

Divide by N, we can get the *unbiased estimation* [1] of r as $\hat{r}$ and its sample variance

$$\hat{r} = \frac{n}{N} \tag{3}$$
$$D(\hat{r}) = \frac{r \cdot (1 - r)}{N} \tag{4}$$

since $E(\hat{r}) = r$.

It turns out that the variance is related to r value. There are things to remember

- The variance is symmetric to 0.5.
- The closer r to 0.5, the larger is the variance.
- The variance decreases when the sample size increases.

Sample Survey

However, in practice, the survey of all population is impossible. Usually, survey in a small group (the number is $M < N$) is available. It will turn out a similar situation

$$\hat{r} = \frac{m}{M} \tag{5}$$
$$D(\hat{r}) = \frac{r \cdot (1 - r)}{M} \tag{6}$$

[1] The unbiased estimation refers the sample mean of the random variable equals to the expectation.

Estimation

Use above analysis, we can say that the estimation of μ value of $R_support$ is easy to compute. But the real question is *How we can trust the estimation?*. Especially in the case of the survey is restricted in *part* of the population.

We analysis the question on both side.

- **In ground-truth end**, we have the equation that

$$\frac{\sigma^2}{N} = \frac{s^2}{N-1}$$

where s^2 refers the sample variance with population of N.

- **In survey end**, we have the variance that

$$s_N^2 = \frac{r(1-r)}{N}$$
$$s_M^2 = \frac{r(1-r)}{M}$$

where s_N^2 and s_M^2 refer the variance of N and M population.

- **Jointly**, since the variance of M is derived from the intrinsic variance σ^2, and population of N. It meets

$$s_M^2 = s_N^2 + s^2$$

Use the joint equation, the M follows

$$M = N \cdot \frac{\frac{r(1-r)}{\sigma^2}}{(N-1) + \frac{r(1-r)}{\sigma^2}}$$

Under certain error edge e, use the equation between e and σ^2

$$e^2 = z^2 \cdot \sigma^2$$

where z is the *Percentile* of normal distribution according to e.

Thus, we have the relationship between e and M.

$$\hat{M} = N \cdot \frac{\frac{z^2 r(1-r)}{e^2}}{(N-1) + \frac{z^2 r(1-r)}{e^2}}$$

The solution is the minimization of *Sample size* to achieve certain degree of confidence defined by e.