Sample Size or Sampling Fraction?
What determines accuracy of a sampling scheme

Rajeeva L. Karandikar
Director
Chennai Mathematical Institute
rlk@cmi.ac.in

Abstract:

The talk will focus on one important issue that props up in the context of sampling and inference based on sample data: how to arrive at a suitable sample size given the objective. Most people think that it is the sampling fraction (sample size as a proportion of the population) that will determine the accuracy of the sample based inference. We will illustrate as to why that is not correct and what is relevant is the sample size and not sampling fraction.

Consider a population of size N. It could be the population of a city, or of a state or a country. The population could also consist of all items manufactured in a given manufacturing facility in a day or a month.

A certain percentage of the population, say $P\%$ possess a characteristics while the rest do not. For example the population of a state has $P\%$ of smokers, or $P\%$ of eligible voters in the country support a political party or $P\%$ of all items manufactured in a given manufacturing facility in a month are defective.

If we are interested in knowing P, and if complete enumeration is not not possible due to paucity of time or resources or some other reason, one resorts to sampling and uses the data from the sample to make a judgement about P.

We will take the simplest scenario , where one uses *Simple Random Sampling* and chooses a sample of size m and determines as to what percentage of the sample possesses the characteristics that we are interested in.

How does one measure accuracy of a sampling scheme? Denoting by $\widehat{P}$ the *estimate* obtained form the sample, we could for example require that

$$Prob(|P - \widehat{P}| \leq 1\%) \geq 99\%$$

or in other words, we would like to be assured with 99% confidence that the estimate does not differ from the true value by more than one percent. More generally, for our chosen values of α, β, we may wish

$$Prob(|P - \widehat{P}| \leq \alpha\%) \geq \beta\%.$$

The question then is, given N, how do we choose sample size to achieve

$$Prob(|P - \widehat{P}| \leq \alpha\%) \geq \beta\% \tag{1}$$

Suppose for a given α, β and $N = 1,00,000$, we have found that $m = 2500$ is sufficient to gurantee (1). If N increases to $N = 4,00,000$, how should n increase- should it go up by factor of 4 or $\sqrt{4} = 2$, *i.e.* should the sample size be 10000 or 5000?

Let us see.

In many situations, we have information on the population and then one can improve upon the *Simple Random Sampling* scheme. For example, to assess $P\%$ of eligible voters in the country support a political party, we know from past data and expert opinions that there is huge variation when one goes from one state to the other. In such a case, one should divide the whole sample in same proportions as population size. This is called *stratified sampling*.

We will consider the case of Simple Random Sampling and illustrate that in this case, it is sample size and not sampling fraction that determines the accuracy of sample based estimate of P.

Suppose that the sample is chosen by repeatedly picking one individual (or item) from the population, with each individual having the same probability $(\frac{1}{N})$ of being picked, at each stage. This assumes that we put back the individual before drawing the next. This is called *Simple Random Sampling With Replacement* (SRSWR).

In this case, writing $X_i = 1$ if the i^{th} chosen individual has the attribute and $X_i = 0$, it follows that $X_1, X_2, \ldots, X_m$ are independent Bernoulli random variables with $P(X_i = 1) = \frac{P}{100} = p$ and thus the number in the sample having the attribute, $S = X_1 + X_2 + \ldots + X_m$ has Binomial distribution with parameters (m, p). Thus, denoting by

$$\hat{p}_m = \frac{X_1 + X_2 + \ldots + X_m}{m}$$

the proportion in the sample, we see that mean of $\hat{p}_m$ is p and variance is $\frac{p(1-p)}{m}$, or the variance is at most $\frac{1}{4m}$.

Thus using Tchebychev's inequality

$$Prob(\ |\hat{p}_m - p| \leq \varepsilon) \geq 1 - \frac{1}{\varepsilon^2}\frac{1}{4m}.$$

Writing $\widehat{P_m} = \hat{p}_m * 100$ - the sample percentage of individuals having the said characteristic we conclude

$$Prob(|P - \widehat{P}| \leq \alpha\%) \geq 1 - \frac{10000}{\alpha^2}\frac{1}{4m}.$$

Taking $\alpha = 2, \beta = 99$, this estimate says that $m = 6250$ would ensure that (1) holds. The population size N did not play any role here, so this is true whatever be N.

Instead of using Tchebychev's inequality, we could use central limit theorem to get a better idea of probabilities for a given m. This was the approach until about 25 years ago when computer power was not available or was expensive. But now we have computing power at our fingertips

We could use **pbinom** command in R to get the probabilities

$$Prob(|P - \widehat{P}| \leq \alpha\%)$$

for various combinations of P, α, m.

For $\alpha = 2$, $m = 6250$ and $P = 60$ this is 99.88% (0.9988) while for $m = 4000$ and $P = 60$ this is just above 99%

We can see that N does not have any bearing on the error probability as it does not enter our formula. This does not quite convince common folks.

Moreover, in practice we will only use *Simple Random Sampling WithOut Replacement* (SRSWOR). Here the samples are no longer iid.

This time the distribution is hypergeometric.

We can see that for $m = 6200$ and $N = 1000000$, the required probability is 99.88%.

And that for $m = 4000$ and $N = 1000000$, the required probability is 99.07%

We could also use simulation to estimate

$$Prob(|P - \widehat{P}| > \alpha\%)$$

This we could do for various combinations of N, m for say $\alpha = 2$ and $P = 60$

Suggest: $N \in \{100000, 400000, 1000000, 4000000\}$ and $m \in \{1000, 2500, 4000, 6250, 8000\}$

However, common folks may still raise doubts about our use of various theorems and derivation of distributions and may not be convinced.

So I have found a nice way of explaining. Using elementary probability taught in class 12, we can claim:

Suppose there are $N1$ red balls and $N2$ white balls in an urn. Total number ways of choosing k red balls out of $N1$ and $(m-k)k$ white balls out of $N2$ is

$$\binom{N1}{k} \times \binom{N2)}{m-k}$$

Total number ways of drawing m balls out of N is $\binom{N}{k}$. Thus, if we draw m balls at random, the probability that we will get exactly k red and $m-k$ white balls is:

$$\frac{\binom{N1}{k} \times \binom{N2}{m-k}}{\binom{N}{k}}$$

As a consequence, from an urn containing $N1 = N * P/100$ red balls and $N2 = N * (100 - P)/100$ white balls, if we draw m balls randomly without replacement (using SRSWOR), the probability that we get between $a1 = m * (P - \alpha)/100$ and $a2 = m * (P + \alpha)/100$ red balls is

$$\sum_{k=a1}^{a2} \frac{\binom{N1}{k} \times \binom{N2}{m-k}}{\binom{N}{k}}$$

Python has built in capabilities to compute large numbers and this probability can be computed exactly even for $N = 1,000,000,000$, $m = 4000$,....

Once again, we suggest Suggest: $N \in \{400000, 1000000, 4000000, 10000000, 100000000, 1000000000\}$ and $m \in \{1000, 2500, 4000, 6250, 8000\}$

This should lay at rest the feeling among skeptics that accuracy depends upon sampling fraction.

Let me now turn to the controversy about EVM-VVPAT that was raked up recently in the weeks leading to the elections to Lok Sabha in April-May 2019.

Let us recall that VVPAT had been introduced to enable courts to

1. assure voters that their vote is correctly recorded
2. to enable recount, in case courts order the same as a result of an election pettition

Let us also note that when supreme court ordered introduction of VVPAT, it did not require EC to have any cross validation of EVM count and VVPAT.

The EC had *suo-moto* announced the policy of randomly choosing one booth per assembly segment and verifying the EVM count and VVPAT count, with the hope that when voters see the matching of these counts, public confidence in EVMs would increase.

However, various political parties continued to attack EVMs and petitions demanding 10% cross validation were filed. In September, the EC had entrusted myself and Prof Abhay Bhatt (from ISI, Delhi) to advice EC on sample size n required, so that if **no mismatches are observed in the n booths chosen randomly between EVM and VVPAt counts, then EC could claim with high confidence that, defective EVMs, if any are negligible**.

To explain the statistical approach adopted by us, let me draw parallel with matching fingerprints of a suspect with that found at a crime scene. The prosecution argument is that, if the suspect was not at the crime scene, the chance that fingerprints would match is extremely low. Thus if finger prints match, the suspect is guilty.

So the opposite of what one wishes to prove with high confidence is taken as *Null Hypothesis* and if under the null hypothesis the observed phenomenon has very low probability, less than α prefixed, we conclude that opposite of *Null Hypothesis* has been proven with high confidence.

We took the desired conclusion as *defective EVM, if any, are less than 2%*. Thus our Null Hypothesis would be **defective EVM are at least 2%** and we are to choose n such that probability of observing **Zero Defectives** in randomly chosen sample of size n is less than α.

We choose α to be the probability of observing a deviation of over 4σ in a normal population, namely $\alpha = 0.00006334248$.

One could compute for each n, the probability of observing **Zero Defectives** in randomly chosen sample of size n - easy to do so today with software. The smallest n for which this probability is smaller than α happens to be 479.

The most interesting part is that this works irrespective of population size. If population is small, say 1000, this number would be smaller, but 479 works for all sample sizes, be it a lakh or ten lakhs or 1 crore.

EC wanted us to give an opinion if their scheme of one EVM per assembly segment would suffice. After obtaining the number of booths in each of the 4125 assembly segments we were able to prove that if we do not observe any defective in the sample drawn then we can conclude with high confidence - 99.993665752% confidence that defective EVMs, if any, are less than 2%

The Supreme Court had suggested that instead of 1 per assembly segment, EC could consider cross validating 5 per assembly segment. We were able to prove that if we do not observe any defective in the sample drawn (of size 20675) then we can conclude with high confidence - 99.993665752% confidence that defective EVMs, if any, are less than $\frac{1}{4}\%$

Let us note that on 23rd May, 5 booths were randomly chosen in each of the 4125 assembly segments and it was found that in all 20675 cases, the EVM count and VVPAT count matched. So we conclude that defective EVMs, if any, are less than $\frac{1}{4}\%$.