EVM sample sizes – a simple guide

Consider the binary indicator E8:05a:

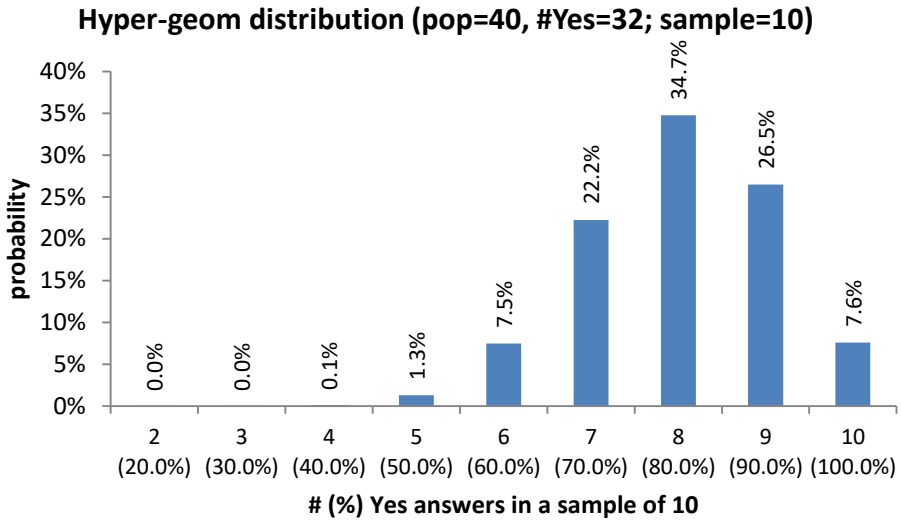
> *Are written instructions on the use of VVMs, such as posters and stickers, available to storekeepers and health workers? [Y, N].*

Suppose there are 40 lowest distribution (LD) vaccine stores in the country of Zizoo, and that for 32 of the 40 stores the answer to E8:05a is Yes. The E8:05a indicator score for all 40 LD vaccine stores is therefore 32/40 = 80%. Suppose now that a random sample of 10 stores is selected from the population of 40 stores. What are the possible E8:05a indicator scores given this sample size, and what are the probabilities of each score?

The probability of getting 8 Yes answers in a sample of 10 is given by[i]:

$$P(8,10;32,40) = \frac{\binom{32}{8}\binom{40-32}{10-8}}{\binom{40}{10}} = 34.7\%$$

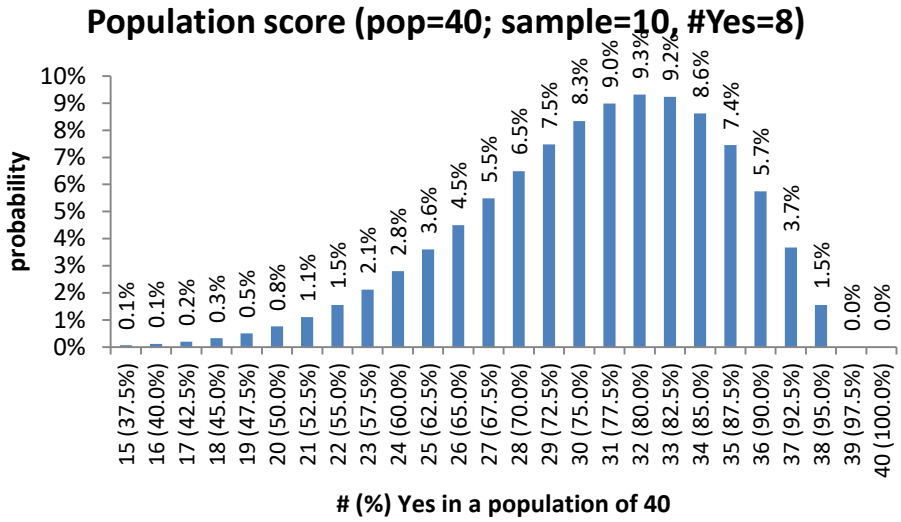
The probability distribution of possible sample scores is shown in the figure below. Such a distribution is called a hyper-geometric distribution.



The chart shows that the sample score lies within ±15% of the population score of 80% with 83.5% confidence.

In the above example, the number of Yes answers in the population of all 40 stores is known, and the number of Yes answers in the sample is to be estimated. During an EVM assessment, the situation is reversed – the number of Yes answers in the sample is known, and the number of Yes answers in the population of all stores is to be estimated.

Reversing the scenario described above, suppose there are 40 LD vaccine stores in the country of Zizoo, from which a sample of 10 are randomly selected, and that for 8 of the 10 selected stores the answer to E8:05a is Yes. The E8:05a indicator score for the 10 selected stores is therefore 8/10 = 80%. Given this sample size and sample score, what are the possible E8:05a indicator scores for all 40 LD stores, and what are the probabilities of each score? The probability distribution of possible scores is shown in the figure below.



The chart shows that the full population score lies within ±15% of the sample score of 80% with 86.8% confidence.

Suppose now that a larger sample size of 20 is chosen, and that for 16 of the 20 selected stores the answer to E8:05a is Yes. The E8:05a indicator score for the 20 selected stores is again 80%. Given this larger sample size, what are the possible E8:05a indicator scores for all stores, and what are the probabilities of each score? The probability distribution of possible scores is shown in the figure below.

With a larger sample size of 20, the chart shows that the full population score lies within ±15% of the sample score of 80% with 98.1% confidence. As expected, when the sample size is increased, the level of confidence in the precision of the estimate is increased.

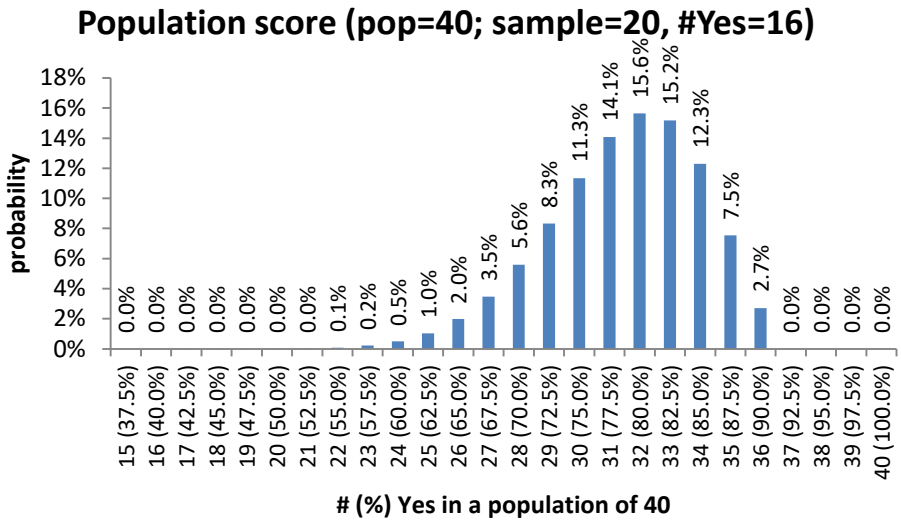


Table 1 shows the sample sizes required for various different population sizes (total number of LD stores). For each population size, the sample size shown is the minimum sample size required to give 85% confidence that the population score is within ±15% of the sample score of 80%.

Total number of LD stores	Minimum sample size
5	4
10	8
25	9
50	11
100	13
250	14
500	15
1000	15

For a given population size, the sample size depends on the desired levels of precision and confidence and on the assumed sample score. The required sample size increases with the desired levels of precision and confidence, and the required sample size increases as the assumed sample score moves away from 50%. Given that the EVM target score is 80%, it is important that scores close to that target be precise, and that their precisions be known. Sample scores closer to 50% need not be so precise. It is for this reason that a sample score of 80% is assumed above.

[i] http://en.wikipedia.org/wiki/Hypergeometric_distribution