Interpreting results of the field surveys using probability calculators

Oleg O Bilukha, Curtis Blanton

Field practitioners in humanitarian settings often face challenges analyzing and interpreting the results of nutrition surveys. Most key variables of interest in field surveys are categorical, i.e., expressed as discrete categories, for example yes/no, or normal/moderate/severe. Examples of categorical variables commonly measured in emergency surveys include prevalence of Global Acute Malnutrition (GAM), stunting, underweight, anemia, coverage of measles immunization and vitamin A distribution programs, and several others.

Some of these variables are “inherently” categorical – for example measles immunization and Vitamin A distribution are measured as yes/no. Other variables, for example anthropometric indicators and anemia, are originally measured as continuous variables (e.g., Hemoglobin concentration for anemia or Z scores for anthropometric indicators), and are converted into categorical variables at the analysis stage using internationally established case definition cutoffs. For example, children 6-59 months of age are classified as stunted if their height for age Z score is <-2, and as non-stunted if Z score is $\geq$-2; the prevalence of stunting is presented as proportion of children classified as stunted among all children in a given sample or population. In this paper, we discuss analysis of key categorical variables measured in field surveys, irrespective of whether they are “inherently” categorical, or have been converted into categorical form from continuous data. The theoretical and practical discourse presented below equally applies to any categorical variable measured as a percentage or proportion of the total.

The most common way to analyze categorical data in the field is to calculate the prevalence estimate and the 95% confidence interval (95% CI) around such estimate. In most cases (unless data analysts have the capacity to perform more advanced statistical analyses), program managers and decision-makers have to rely on these three numbers -- prevalence estimate, lower confidence limit, and upper confidence limit -- to interpret the results and make programmatic decisions.

The key underlying idea in using the estimated prevalence from a representative sample survey is that a (often relatively small) fraction or sample of the population can provide a reliable estimate of the *true population prevalence*. For example, the prevalence of GAM measured from a survey of 500 children (*sample prevalence estimate*) would be sufficiently close to the prevalence of GAM in all 100,000 children in the surveyed population (*true population prevalence*). Note that we cannot measure all 100,000 children, and therefore we will never know for sure what the *true population prevalence* is, but instead rely on a *sample prevalence estimate* and the 95% CI limits to provide a range where the true population prevalence is most likely to lie. In the surveys where collected data are valid and representative, there is a 95% probability that the true population prevalence lies between the lower and the upper limits of the 95% CI (note that there is still a 5% probability that the true population prevalence lies outside of the 95% CI limits). For example, in the survey where GAM prevalence estimate is 12% and the 95% CI limits are 8% and 16%, there is 95% chance that the true population prevalence of GAM lies somewhere between 8% and 16%.

The key goal when analyzing and interpreting categorical survey data is often to infer not only how high or low the *true population prevalence* is likely to be, but also how likely is it to exceed the pre-determined action thresholds (e.g., 5%, 10%, 15% for GAM;[1] 20% and 40% for anemia;[2] etc.) The chance of the true population prevalence falling within a given range is described by the area under the probability distribution curve. For example, Figure 1 presents a binomial probability distribution curve for the prevalence estimate of GAM from the survey example above. As can be seen, 95% of the area under the distribution curve falls between lower and upper 95% CI limits, whereas 2.5% of the area under the curve falls below the lower 95% CI limits, and 2.5% of the area falls above the upper 95% CI limit. Therefore there is a 2.5% chance that the true population value is below the lower 95% CI limit, and 2.5% chance that the true population value would be higher than the upper 95% CI limit. Similarly, from Figure 2, since 50% of the area under the distribution curve lies below the survey prevalence estimate, and 50% lies above, we can conclude that there is an equal chance that the true population prevalence would be below or above the survey prevalence estimate (in this example, the true population prevalence of GAM is equally likely to be below or above 12%).

Figure 1. Probabilities associated with 95% Confidence Interval

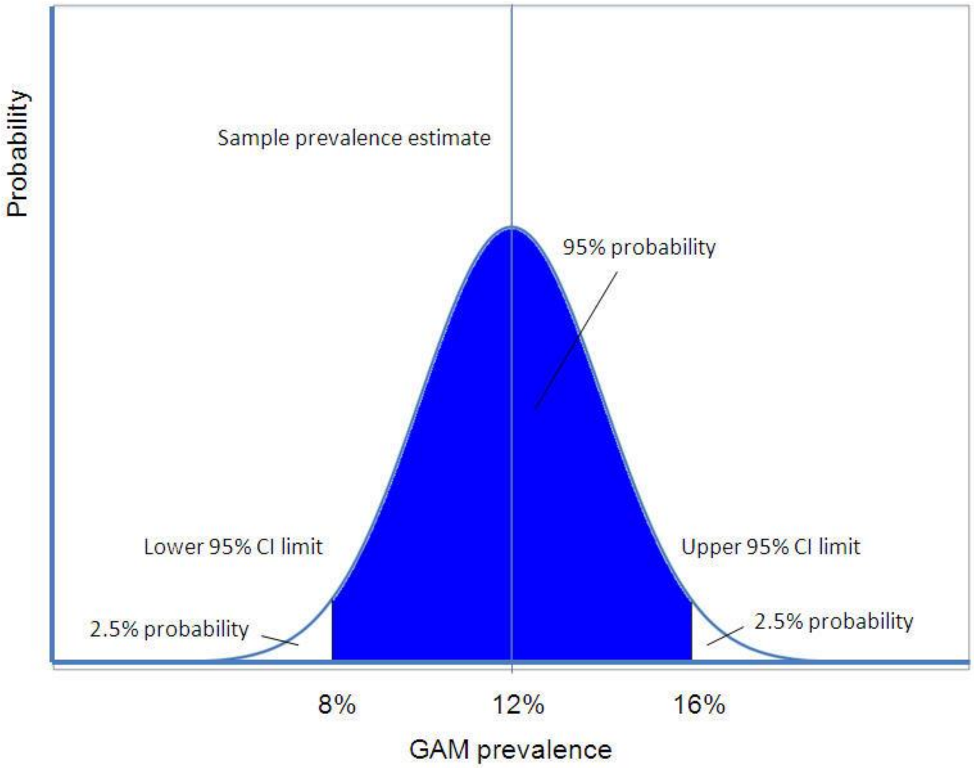
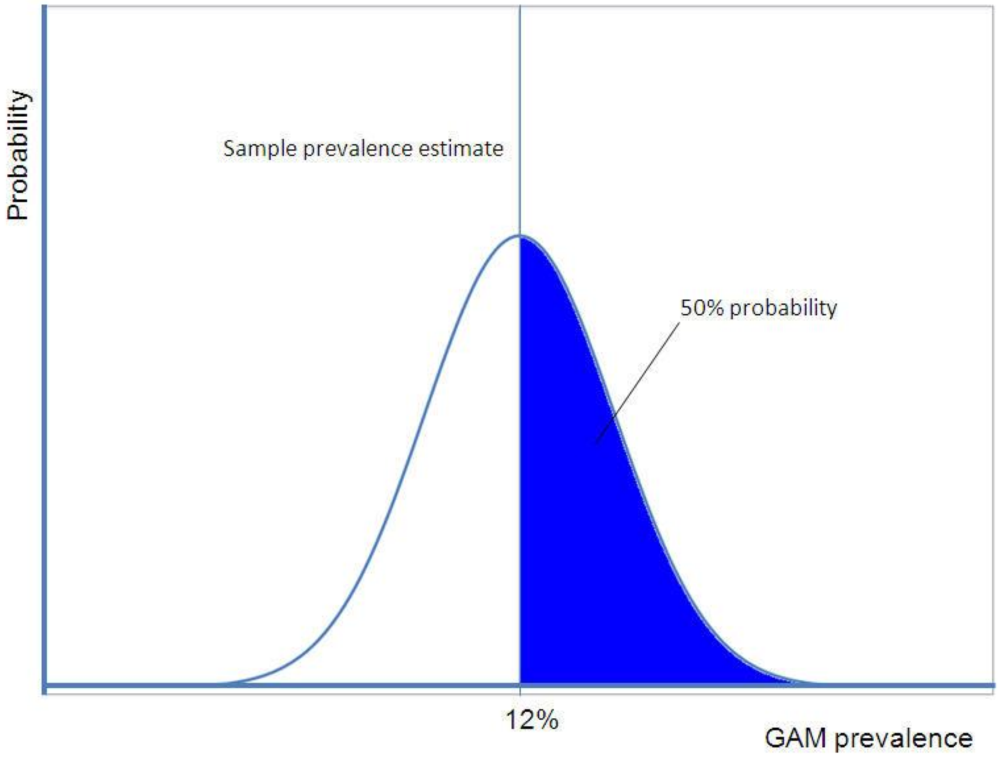


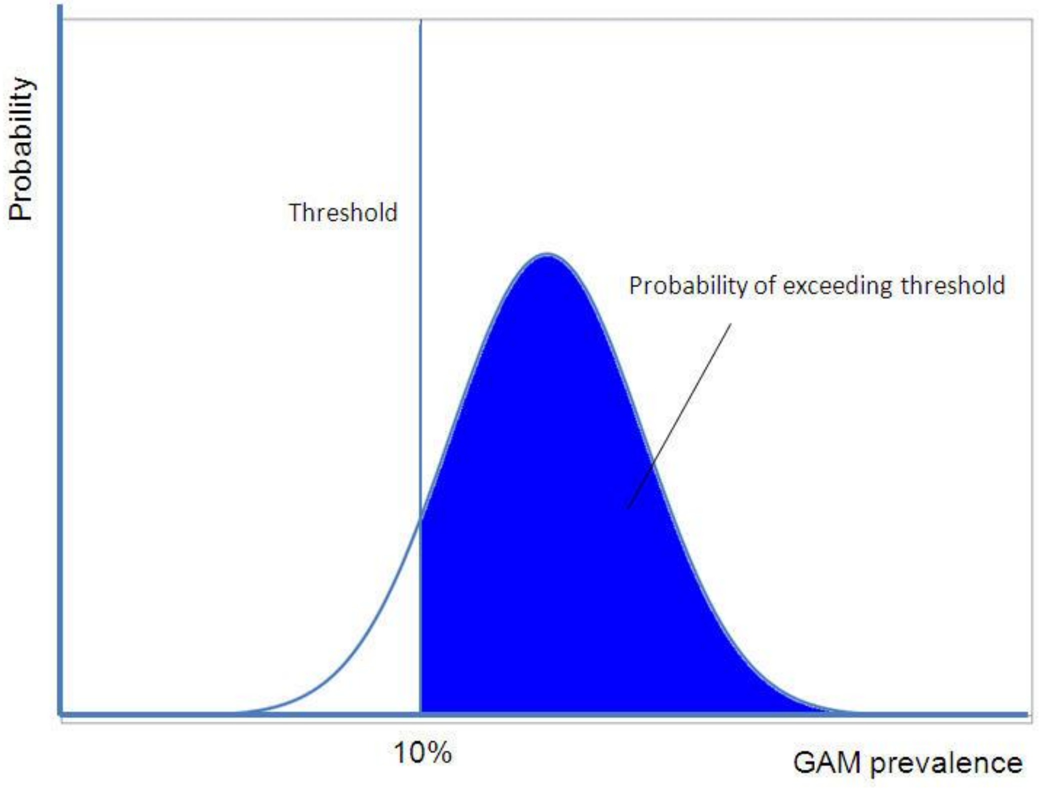
Figure 2. Probabilities associated with sample prevalence estimate



When we look at the practices presently used in the field, the most common way of classifying GAM (or other indicators) relative to the thresholds is based solely on the magnitude of the survey prevalence estimate (e.g., if the GAM prevalence observed in the survey exceeds the threshold, then the area is declared above the threshold, and vice-versa). From a statistical perspective, this means that GAM is declared above the threshold when statistical probability of the true population value of GAM exceeding the threshold is above 50%. One drawback of this approach is that the width of the confidence interval becomes virtually irrelevant; it may be, in fact, often ignored in summarizing the data for decision-making. Another question is whether 50% constitutes sufficient “risk” or “confidence” to make programmatic decisions.

When comparing survey results to pre-determined thresholds, the primary interest is to estimate the probability, or “risk,” that the true population prevalence exceeds the threshold. The higher the “risk,” the more seriously decision-makers would need to consider implementing appropriate interventions. The probability of the true population prevalence exceeding the threshold is described by the area under the distribution curve that falls above the threshold, as depicted in Figure 3. Using our previous survey example, the area under the curve represents the probability of the true population prevalence to exceed the 10% threshold.

Figure 3. Probability of exceeding the threshold



To provide additional information for decision-making, we developed a "threshold" probability calculator that provides the estimated probability of the true population prevalence exceeding the threshold. We used a one-sided t-test for proportions, where the alternative hypothesis tested is that the true population prevalence is lower than the threshold. P-value for this test provides an estimated probability (or "risk") that the true population prevalence exceeds the threshold.[3,4]

The calculator is in a spreadsheet format, where the user needs to enter some summary survey statistics to obtain the probabilities of exceeding the thresholds. There are 3 versions of the calculator (included as separate spreadsheets on the Excel file):

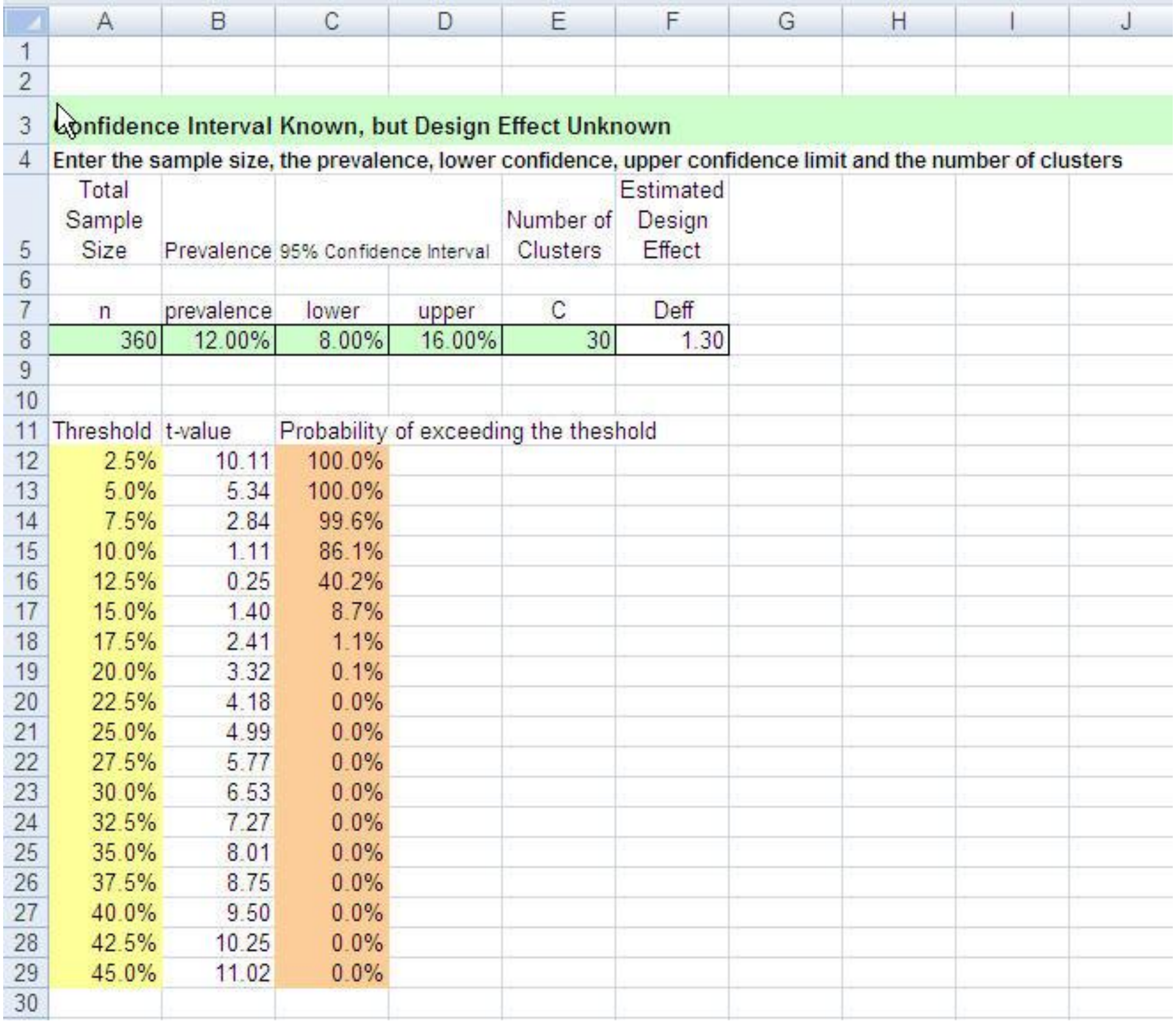
1. To use for cluster survey designs, when the design effect (DEFF) for the indicator is known. In this case, the user needs to enter total survey sample size, the number of clusters, survey prevalence estimate, and the DEFF

2. To use for cluster survey designs, when DEFF for the indicator is not known. In this case, the user needs to enter total survey sample size, the number of clusters, survey prevalence estimate, and the upper and lower 95% CI limits for this estimate

3. To use in simple or systematic sample surveys. In this case, the user needs only to enter total survey sample size and survey prevalence estimate.

Figure 4 provides the screenshot of the calculator. The information mentioned above is entered in the green cells. The thresholds for which the probabilities are provided are in the yellow column. These thresholds can be defined/changed by the user. The probabilities of the true population value exceeding the threshold are calculated automatically and displayed in the orange column. Figure 4 provides an example of the survey that we used in discussions above (GAM prevalence estimate of 12% and the 95% CI limits 8% to 16%), assuming that this was a cluster survey with 30 clusters and a total sample size of 360 children.

Figure 4. “Threshold” calculator



Confidence Interval Known, but Design Effect Unknown

Enter the sample size, the prevalence, lower confidence, upper confidence limit and the number of clusters

Total Sample Size	Prevalence	95% Confidence Interval		Number of Clusters	Estimated Design Effect
n	prevalence	lower	upper	C	Deff
360	12.00%	8.00%	16.00%	30	1.30

Threshold	t-value	Probability of exceeding the theshold
2.5%	10.11	100.0%
5.0%	5.34	100.0%
7.5%	2.84	99.6%
10.0%	1.11	86.1%
12.5%	0.25	40.2%
15.0%	1.40	8.7%
17.5%	2.41	1.1%
20.0%	3.32	0.1%
22.5%	4.18	0.0%
25.0%	4.99	0.0%
27.5%	5.77	0.0%
30.0%	6.53	0.0%
32.5%	7.27	0.0%
35.0%	8.01	0.0%
37.5%	8.75	0.0%
40.0%	9.50	0.0%
42.5%	10.25	0.0%
45.0%	11.02	0.0%

From the values in the orange column on Figure 4 we can see that in this survey area, the probability of the true population value of GAM to exceed 5% threshold is close to 100%, and probabilities of exceeding 10%, 15% and 20% thresholds are 86%, 9% and 0.1%, respectively. This provides much richer information on population “risk” for decision makers, compared with information based solely on the prevalence and confidence interval limits described above. For example, it tells the user that it is quite likely (86% probability) that the true value of GAM exceeds the 10% threshold, and quite unlikely (9% probability) that the true value of GAM exceeds the 15% threshold. We believe that this information directly quantifying the "risk" of the true population prevalence exceeding the threshold, combined with other contextual information on risk and protective factors should prove useful for decision-making.

Note that we do not intend to discuss what level of "risk" (25%, 50%, 95% or other) is high enough to be taken "seriously" and trigger action. We believe that these decisions should be context-specific, and action should be considered taking into account both the statistical "risk" estimated from survey data as well as other existing and potential risk factors.[5] Note also that we do not necessarily endorse the appropriateness of currently used action thresholds for various indicators, or the concept of making programmatic decisions based on comparing the observed prevalence to pre-existing thresholds. We only provide a convenient statistical tool for those field practitioners who feel compelled to conduct these types of analyses.

The calculator presented on Figure 4 can be used for any categorical variable for which results are expressed as a proportion (or percentage) of the total – for example, for prevalence of anemia, immunization coverage, stunting, wasting, etc. As mentioned, the thresholds can be changed as necessary for a given indicator. For example, it is possible to test what is the probability that measles immunization coverage exceeds a minimum acceptable level, or whether anemia prevalence exceeds programmatic action threshold that calls for blanket iron supplementation, etc.

Another challenge for field practitioners is presented when the situation requires assessing significance of the difference between two survey results. For example, consider testing the difference between the surveys conducted in the same area in 2 different seasons or in 2 different years; or testing the differences between the results obtained from the surveys in 2 neighboring districts or livelihood zones. In these cases, field practitioners often use the "overlapping confidence intervals test" – i.e., if the 95% CI limits around the estimates from 2 surveys do overlap, the results are declared not statistically different, and if confidence limits do not overlap, the results are considered statistically different. The problem is, that in many instances when confidence intervals do overlap slightly, results may still be significant at 95% confidence level. This is especially true if a one-sided test can be used as discussed below.

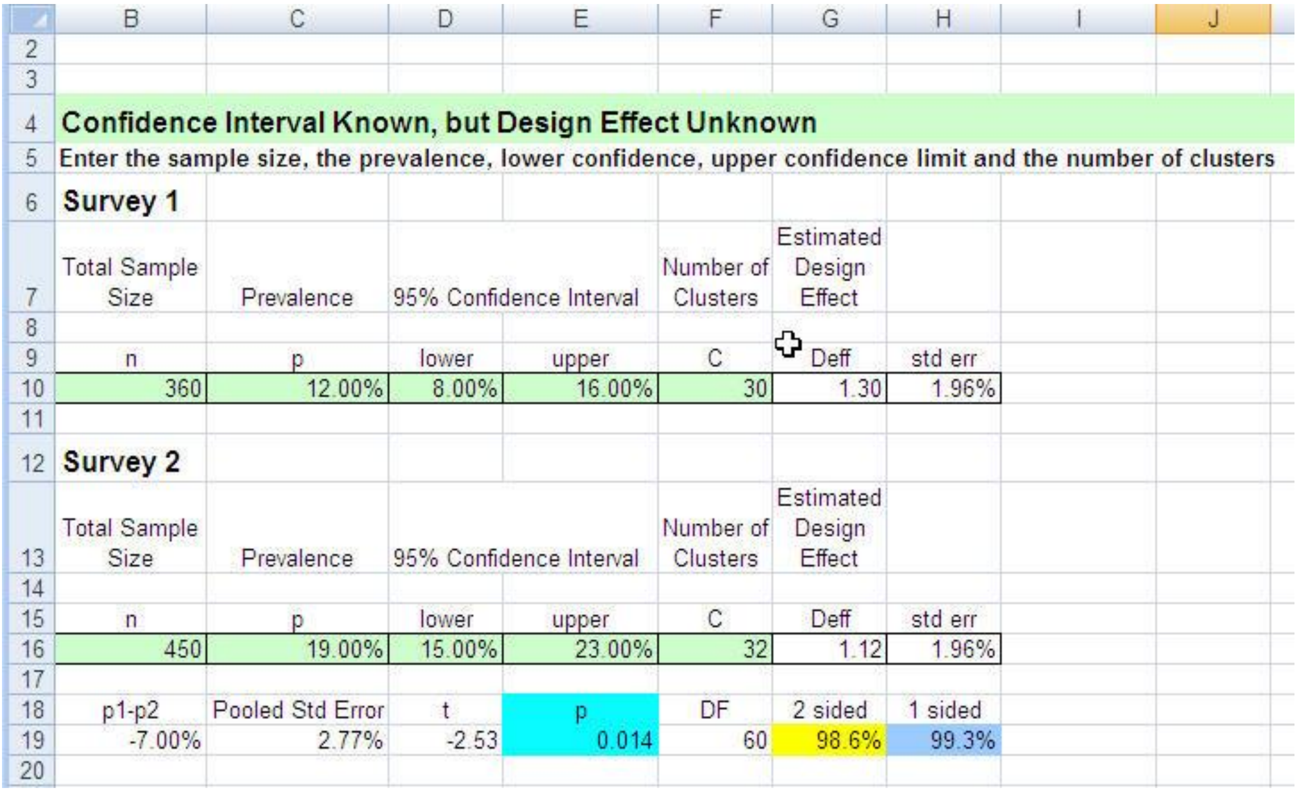
To assist field practitioners in these situations, we developed a "two-survey" calculator for testing the statistical significance of the difference between the estimates from 2 surveys (or from 2 strata of the same survey). The statistics in this calculator are based on a t-test for the difference between 2 proportions, testing an alternative hypothesis that the true population values in the 2 surveys are different from each other.[6,7] The two-tailed probability that the true population values are different from each other is calculated as 1-p, where p is a p-value of the above t-test for 2 proportions. The calculator provides both 1-tailed and 2-tailed probabilities.

Similarly to the threshold" calculator, the "two-survey" calculator is also available in Excel format and has 3 spreadsheets:

1. For cluster survey designs where prevalence estimates and DEFF in both surveys are known
2. For cluster survey designs where prevalence estimates are known but DEFF are unknown
3. For simple or systematic random surveys

The information that users need to enter for each of the surveys is the same as in the "threshold" calculator.

Figure 5. “Two-survey” calculator



	B	C	D	E	F	G	H	I	J
2									
3									
4	**Confidence Interval Known, but Design Effect Unknown**								
5	**Enter the sample size, the prevalence, lower confidence, upper confidence limit and the number of clusters**								
6	**Survey 1**								
7	Total Sample Size	Prevalence	95% Confidence Interval		Number of Clusters	Estimated Design Effect			
8									
9	n	p	lower	upper	C	Deff	std err		
10	360	12.00%	8.00%	16.00%	30	1.30	1.96%		
11									
12	**Survey 2**								
13	Total Sample Size	Prevalence	95% Confidence Interval		Number of Clusters	Estimated Design Effect			
14									
15	n	p	lower	upper	C	Deff	std err		
16	450	19.00%	15.00%	23.00%	32	1.12	1.96%		
17									
18	p1-p2	Pooled Std Error	t	p	DF	2 sided	1 sided		
19	-7.00%	2.77%	-2.53	0.014	60	98.6%	99.3%		
20									

Figure 5 presents a screenshot of the “two-survey” calculator. Users enter information in the green cells, the p-value is presented in the turquoise colored cell, the 2-tailed probability is in the yellow cell, and the one-tailed probability is in the blue cell.

Consider comparing GAM prevalence from the 2 surveys conducted in neighboring districts A and B (Figure 8). District A results are the ones we used as an example in a “threshold” calculator, and district B results are as follows: GAM prevalence of 19%, 95%CI from 15% to 23%, sample size 450, 32 clusters. Note that the 95% CI for the 2 surveys overlap (8%-16% in survey A and 15%-23% in survey B), so by the “overlapping confidence intervals test” the difference between two surveys would be declared non-significant. From the output in Figure 5, however, we see that the p-value for the 2-tailed test (p=0.014) is significant at 0.05 level, and the 2-tailed probability is 98.6%, meaning that there is about 98.6% statistical probability that the true prevalence of GAM in districts A and B are different from each other.

So, when should we use 1-tailed versus 2-tailed test? For most comparisons between 2 surveys a 2-tailed test would be an appropriate test to use. It is more conservative of the two, and does not depend on the *a priori* hypotheses. The 1-tailed test is more powerful (it always returns a higher probability that 2 surveys differ from each other), but must be used cautiously and only in specific situations. Generally, we can use a 1-tailed test if we have an *a priori* hypothesis that one population’s prevalence is higher than the other, and can clearly justify our thinking. For example, we could use a 1-tailed test in our example above if *before* doing surveys in districts A and B we could *publicly* declare that we expect GAM to be higher in District B, and could explain why we expect that (e.g., because blanket supplementary feeding and general food distribution are implemented in District A and not B, or because District B and not District A experienced drought and had poor harvest, etc.) Note that if our *a priori* guess turns out to be incorrect (e.g., we expected GAM to be higher in District A, and the surveys showed a higher GAM in District B), we cannot use a 1-tailed test.

As was the case with the “threshold” calculator, the “two-survey” calculator can also be used for any categorical variable for which results are expressed as a proportion (or percentage) of the total – for example, for prevalence of anemia, immunization coverage, stunting, wasting, etc.

In conclusion, we wanted to emphasize that analyses performed by these calculators can also be performed using any common statistical software, like SPSS, SAS or STATA. We propose them solely for their convenience, realizing that field practitioners often do not have advanced skills in data management and analysis, or do not have access to statistical software that require expensive licensing rights.

The calculators described in this paper are available from the website of the International Emergency and Refugee Health Branch, CDC: http://www.cdc.gov/nceh/ierh/

We look forward to a feedback from field practitioners on the use of these tools. Please send your questions, comments or suggestions to Dr. Oleg Bilukha: obilukha1@cdc.gov

References

1. World health Organization: Management of Nutrition in Major Emergencies. Geneva: WHO; 2000.
2. World Health Organization: Iron Deficiency Anemia. Assessment, Prevention and Control. Geneva: WHO; 2001.
3. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Fam Pract 2000, 17:192-6.
4. Fleiss JL; Levin B; Paik MC. Statistical Methods for Rates and Proportions, 3rd ed. New York: John Wiley & Sons; 2003.
5. Bilukha OO, Blanton C. Interpreting results of cluster surveys in emergency settings: is the LQAS test the best option? Emerg Themes Epidemiol 2008, 5:25. http://www.ete-online.com/content/5/1/25
6. Murray DM: Design and Analysis of Group Randomized Trials. New York: Oxford University Press; 1998.
7. Donner A, Klar N. Design and Analysis of Cluster Randomization Trials in Health Research. New York: Hodder Arnold; 2000.