Statistical Consulting on Human Rights and Humanitarian Projects

Susan Hinkins
NORC at the University of Chicago, Chicago, United States – Hinkins@mcn.net

Abstract

In any project involving statistical procedures, the statistician has a professional responsibility to ensure that the statistical questions are well defined, the data are collected and edited responsibly, with full documentation, that the statistical procedures are used correctly and that the results are accurately represented in the final report. This last responsibility is critical. Projects associated with human rights or humanitarian concerns have these same requirements and the same issues. Resources are often limited and there is a desire to have strong results quickly. In this presentation, an example is provided where a particular statistical technique was simplified for general use, and in some cases the purpose and the correct statement of conclusions has been misunderstood, with potentially serious results. Human rights and humanitarian organizations that rely on statistical tests should be encouraged to develop an ongoing professional relationship with a statistical group and maintain a statistician on staff to review all materials.

Keywords: LQAS; hypothesis tests.

1. Introduction

One of the advantages in working as a consulting statistician is the opportunity to team with individuals in a variety of disciplines to investigate a variety of issues. The topic for this session is to illustrate statistical consulting on humanitarian and human rights projects. This area provides a wide range of problems where statisticians can provide value. Human rights and humanitarian projects are no different in terms of the statistician's responsibilities and role, specifically the following two broad roles:

1) Educating clients on the roles and responsibilities of the statistician.
2) Providing principled findings with adequate documentation.

Every consulting statistician has had the experience of being asked to help with a statistical question or calculation which 'will not take much time.' There is no such thing. In such cases, the statistician must educate the client on the inherent responsibilities of the statistician. Ethically, we cannot provide statistical calculations and conclusions about data without knowing why and how the data were collected and edited, or otherwise manipulated. The statistical team must be included in all aspects of the data collection, cleaning, and editing. There is often a sense of needing 'statistical significance' without a clear understanding of what this might mean in the context of the problem. The client must also understand that the statistician cannot guarantee to find 'significant effects' nor 'prove' a particular point. The statistician's responsibility is to determine what the data contain, i.e. what can and what cannot be determined from the data.

To the second point, the statistician must play a key role in writing and approving the wording of the findings and the final report. This includes explaining the technical nuances and reviewing all reports for the accuracy of the statistical statements. Statements of statistical conclusions require very precise language and it is not uncommon for clients to overstate the statistical results, usually due to misunderstandings about the limitations of the methodology. One of the most important sections of a final report is the "Limitations" section where issues such as coverage, nonresponse and model assumptions are discussed.

The most successful teams build up a mutual understanding of the project– the ultimate goal, the assumptions being made, the methods and issues in data collection, etc. The statistician needs to be a team member. Therefore human rights or humanitarian organizations (HROs) who 'routinely' use statistical analysis on an ongoing basis, should be encouraged to hire a statistician on staff, who can then determine when additional expertise may be needed.

The example I will discuss today is an example where it appears that in some cases, the lack of consistent statistical assistance has led to a critical misunderstanding of a statistical procedure. In this example, a methodology was developed under specific assumptions, to address a specific issue, but as the description of the test was 'simplified' it became more likely to be misunderstood and then misapplied.

2. Lot Quality Assurance Sampling

The Lot Quality Assurance Sampling (LQAS) test is routinely used to monitor the progress being made in providing health interventions in developing countries. I became aware of its widespread use when the ASA Committee on Scientific Freedom and Human Rights was asked to review the USAID training manuals for this technique. Subsequently I reviewed the use of this methodology via Statistics without Borders (SWB) projects where the test was used for evaluating the effectiveness of health interventions, such as vaccination programs and childhood health care.

First, I want to emphasize that the LQAS test is a very useful technique, as described in the book by Valadez [1], and it is often used to great advantage. The technique was initially applied in an industrial context to test whether a 'lot' was acceptable or unacceptable, defined in terms of the number of defects. In the application to assessing health programs in developing countries, a 'lot' is typically defined in terms of individuals in a geographic region who should have received a treatment such as a vaccination or education or food from members of a particular organizational unit (supervisory area). The LQAS test was developed as a management technique to identify areas or teams where the process was not being successful and where additional resources were needed in order to ensure ultimate success. "The basic aim of the method is to identify substandard practices which might produce low quality service delivery." [2] In addition, the test should also have low probability of mistakenly identifying a 'successful' team as requiring remediation as that would waste resources.

Two parameters are needed to define these two objectives:

- The ultimate goal for success, in terms of the percentage of individuals treated
- The definition of 'failure' at the intermediate point where the process is being tested.

Valadez describes an example. "The Director General of Health selected an 80%:50% triage system of this assessment. He expected that at least 80% of the CHWs[3] would perform adequately the tasks of each subsystem. Subsystems in which 50% or fewer of the CHWs performed up to the standard were considered priorities for national reform."

There are two risks to be considered in this problem. The consumer risk is the risk that the test fails to identify a 'failing' supervisory area, and the provider risk is the risk that the test incorrectly identifies a successful area as requiring remediation. The test is devised so that the consumer risk and the provider risk are both small (typically less than 10%) and approximately equal.

[1] Valadez, Joseph J. (1991), *Assessing Child Survival Programs in Developing Countries*, Harvard University Press.

[2] Valadez, Joseph J. (1991), p. 129

[3] Community Health Worker

Once these parameters are determined, e.g. the goal for success is 80% and failure is defined at 50%, one can determine the smallest sample size where a test can be defined having the desired level of risk. In order to determine the appropriate test, Valadez provides cumulative *binomial probability* calculations for a range of sample sizes. For each sample size, n, Valadez provides an entire page of cumulative probabilities, where the table column is associated with the true value of p, the proportion of success, and each row corresponds to a value d, the number of defects observed. The cell value is the cumulative binomial probability of observing at most d defects given the value of p and the sample size n. In this example, if the desired risk is 10%, the smallest sample size is n=19. Exhibit 1 shows a portion of the tabled values for n=19.

Exhibit 1. Examples[4] of Cumulative Probabilities for Sample Size n=19

d	**Values of True p: Proportion of success**								
	0.45	***0.50***	**0.55**	**0.60**	**0.65**	**0.70**	**0.75**	***0.80***	**0.85**
5	0.011	0.032	0.078	0.163	0.297	0.474	0.668	0.837	0.946
6	0.034	0.084	0.173	0.308	0.481	0.666	0.825	0.932	0.984
7	0.087	0.180	0.317	0.488	0.666	0.818	0.923	0.977	0.996
8	0.184	0.324	0.494	0.667	0.815	0.916	0.971	0.993	0.999

The table indicates that for the 80%:50% triage system described earlier, if we want each risk to be approximately 0.10, then the appropriate test with n=19 would be to identify as ‘acceptable’ any lot where d, the number of failures, is 6 or less; no additional resources would be expended on a ‘lot’ if there were 6 or fewer failures. This test has approximately equal consumer and provider risk, each slightly less than 0.10. That is:

1. If the true success rate is 50%, the probability of identifying the lot as acceptable is 0.08 (Consumer risk)
2. The probability of rejecting the lot given that the true success rate is 80% is 1-0.932= 0.07 (Provider risk)

Valadez’s book discusses the test in terms of the number of failures which will trigger rejecting a lot as being acceptable and states[5] “the utility of LQAS is its ability to identify rapidly deficient health facilities within a larger area.” This is a very useful technique and Valadaz clearly describes its purpose and limitations. He makes a point, repeatedly, of the need to consider and determine **both** parameters of interest (80%:50% in this case).

3. An Example of Misunderstanding of LQAS

In various reports based on the LQAS technique and in training material we reviewed[6], a simplification of the tables may have led to potentially serious misunderstandings regarding the purpose of this test. Only one of the two necessary parameters remains in the description and it appears that, in some cases, belief has developed that the test in the previous example can be used to test whether the desired success rate of 80% has been achieved. This misunderstanding could lead to serious consequences.

First, the table (and the test) is often converted from a test based on the maximum number of defects to a test based on the minimum number of successes. This is a reasonable adjustment, but the use of the number of successes may subconsciously translate the thought process to view this as a test of having

[4] Valadez, Joseph J. (1991), Appendix p. 189

[5] Valadez, Joseph J. (1991), p. 93

[6] Some examples include a set of slides used for training by CORE Monitoring and Evaluation Working Group, provided in the references, and a USAID/ENGINE Project mid-Term Report.

reached the desired level of success, rather than identifying lots that require remediation or special attention.

The critical issue is that only one of the two parameters is identified in the table; the key parameter defining 'unacceptable' is absent. This allows the reduction from many pages of tables to one table for many sample sizes. The table provides decision rules for various sample sizes and for a range of upper bounds, with essentially no discussion of the second parameter needed to define the test (50% in the previous example).

Exhibit 2 reproduces some of the columns and rows for three sample sizes. In the LQAS test, if the ultimate goal for success is 80%, then n=19 is the smallest possible sample size.

Exhibit 2. Example of 'Simplified' Table
Decision Rule in Terms of Minimum # of Successes

	Ultimate Goal of Success							
Sample Size	55%	60%	65%	70%	75%	**80%**	85%	90%
19	8	9	10	11	12	**13**	14	15
20	8	9	11	12	13	14	15	16
21	9	10	11	12	13	14	16	17

The shaded cells for n=20 indicate choices where the "Alpha and Beta are > 10%" and therefore these are not options as the risk of error is greater than 10%. This mention of both Alpha and Beta is the only indication of the second parameter of interest.

The same test described above, based on the 80%:50% decision rule, can be identified under the column for 80%, namely that the area will be deemed as "not a priority for additional resources" if there are at least 13 successes out of 19. However there is nothing in the table to remind one that the test is performed to identify substandard areas which are defined as those with a success rate of **50%** or less.

In the training material, the resulting action based on the test is described correctly in the sense that if fewer than 13 successes are observed, then the supervisory area is classified as requiring intervention. Otherwise the supervisory area is considered 'adequate'. However the acceptance is described as "no statistical evidence that performance is < 80%." This wording is misleading.

Unfortunately, in some cases it appears that obtaining at least 13 successes out of 19 was mistakenly interpreted as confidence that the performance is at least 80%, when in fact it is the lower parameter of a 50% success rate that should be used in such a statement. In other reports, there are indications of similar misconceptions that an observed 68% success rate (13/19) in a sample of size n=19, provides confidence that the goal of 80% success has been achieved.

In summary, the progression of statements of conclusion can be described as the initial, correct statement

> "we are confident that we will not reject the lot if the true success rate is 80%."

This was then restated as

> "no strong statistical evidence that performance is < 80%."

And then, in some cases, it became

> "acceptance that the performance rate is at least 80%"

By simplifying the description of the methodology, a critical parameter has been eliminated from the discussion.

In the example above, the LQAS test has essentially been reconfigured and thought of as a standard hypothesis test. I would argue that it is a bad idea to attempt to think of the LQAS test as a hypothesis test. The problem is that it is very important how you frame the hypothesis and it is important to consider both Type I error and Type II error, or equivalently the power of the test. Hypothesis tests are typically set up so that the null is the status quo and we require significant evidence in order to reject it. Ideally one should also define a material difference and consider the power of the test to distinguish a material difference. But in practice often only the Type I error is considered.

Using the same example, there are two choices for how it could be reconfigured. The reasonable one would be to make the null hypothesis H_0: $p \leq 0.50$ where p is the true proportion of successes, e.g. true proportion of children vaccinated. Exhibit 3 summarizes this choice.

Exhibit 3. LQAS Reconfigured as testing H_0: $p \leq 0.50$
Reject H_0 if $X > 13$ where X is the number of successes out of 19

Errors	Cost of Making the Error	Probability of Error
I: Reject H_0 when $p \leq 0.50$	Fail to add resources to a 'failing' area	Pr(Type I error) ≤ 0.08
II: Accept H_0 when $p > 0.50$	Add resources to project which is not 'failing.'	Pr(Type II error| p=0.80) =0.07

With such a small sample size, one does not expect much power and, as shown in Exhibit 4, there is only reasonable power for alternative values of p which are 0.70 or greater.

Exhibit 4. Properties of the Test for H_0: $p \leq 0.50$

Property	Alternative Values of p, greater than 0.5					
	0.55	0.60	0.65	0.70	0.75	0.80
Pr(Type II Error)	0.8	0.7	0.5	0.3	0.2	0.1
Power to Detect	0.2	0.3	0.5	0.7	0.8	0.9

The fact that there is very little power to detect when the true value is p=.60 may be reasonable. That is, using this test, when the true p=.60, we can expect to 'add resources' 70% of the time. In this context, it seems very reasonable to accept this likelihood of an error rather than invest in larger sample size.

However, consider the similar discussion if the test is viewed as testing the null hypothesis that the true success rate is at least 80%.

Exhibit 5. LQAS Reconfigured as testing H_0: $p \geq 0.80$
Reject H_0 if $X \leq 13$ where X is the number of successes out of 19

Errors	Cost of Making the Error	Probability of Error
I: Reject H_0 when $p \geq 0.80$	Fail to identify a successful project	Pr(type I error) ≤ 0.07
II. Accept H_0 when $p < 0.80$	Fail to provide service to desired proportion of the population.	Pr(Type II error| p=0.50) = 0.08

The sample size is too small to identify when the true success rate is noticeably smaller, as shown in Exhibit 6. This is not the null hypothesis that should be tested; success should not be assumed to be the status quo. The cost of making a Type II error is now the cost of not identifying a failure to deliver aid, and in most cases, one would want much more power to detect a lack of success than can be provided with a sample size of 19. For example, is it appropriate to use a test that has a 50% chance of accepting the hypothesis of 0.80 success when the true success rate is only 0.65?

Exhibit 6. Properties of the Test for H_0: $p \geq 0.80$

Property	Alternative Values of p, p < 0.80					
	0.50	0.55	0.60	0.65	0.70	0.75
Pr(Type II error| p)	0.1	0.2	0.3	0.5	0.7	0.8
Power to Detect	0.9	0.8	0.7	0.5	0.3	0.2

The LQAS test was not intended to be used in this manner and this particular misunderstanding and misapplication of the LQAS test has been pointed out by others, for example Rhoda et al (2010). This misunderstanding could have serious consequences when, for example, it is believed that the test indicates that 80% of the individuals have been vaccinated when in fact they should only have confidence that at least 50% have been vaccinated. The simplification removed a key parameter from the discussion and from the statement of results, and may have led to a belief that a sample of size 19 provides sufficient power to test an important hypothesis.

This is an example, I believe, of the potential difficulties that can arise when the statistician is removed from the team. The wording of statistical results can be cumbersome and it is not a rare occurrence for the client to reword such statements in an effort to make them more easily understood. It is the statistician's responsibility to review all such reports and ensure that the wording does not misrepresent the test or the results.

4. Summary

Statistics is a tool which can be of benefit in many areas, including the measurement or evaluation of metrics associated with human rights issues or humanitarian projects. Like all other areas where statistical inference is used, it is critical that the statistician be fully involved in the project and not asked to 'drop in' at the end to make a few calculations. It is also vital that the statisticians stay involved with such projects to ensure that the methodology continues to evolve and improve as needed. Human rights or humanitarian organizations that use statistical analysis on an ongoing basis should be encouraged to hire a statistician on staff, who can then determine when additional expertise may be needed.

References

CORE Monitoring and Evaluation Working Group (February 2006). "LQAS Online Series. Lecture #1. Introduction to Lot Quality Assurance Sampling. Basic Principles." http://207.226.255.123/conf_reg/LQAS_Lecture_1.pdf. Accessed on 5-24-2010

CORE Monitoring and Evaluation Working Group (March 2006). "LQAS Online Series. Lecture #5. Using LQAS for Monitoring."

Rhoda, Dale A., Soledad A. Fernandez, David J. Fitch, and Stanley Lemeshow (2010), LQAS: User Beware. *International Journal of Epidemiology*; **39**:60–68.

Valadez, Joseph J. (1991), *Assessing Child Survival Programs in Developing Countries*, Harvard University Press.