2019

A Case for a Risk-Based Sampling Approach to Manual Accessibility Testing

MARK TURNER, PHD AND ARTHAK ADHIKARI

A Case for a Risk-Based Sampling Approach to Manual Accessibility Testing

Lead Author: Mark Turner, Ph.D. (CEO)
Contributing Author: Arthak Adhikari (Research Analyst)

Optimal Solutions Group, LLC

University of Maryland Discovery District
5825 University Research Court, Suite 2800
College Park, MD 20740
United States
Email: info@optimalsolutionsgroup.com

Abstract

Automated testing can only assess up to 30-35 percent of the WCAG standards. To ensure that web products conform, it is necessary to conduct manual accessibility testing. Manual WCAG testing is labor intensive and costly. To ensure that organizations are able to routinely monitor web products' WCAG compliance that are statistically valid, Optimal Solutions Group ("Optimal") has designed an iAccessible module to generate a risk-based stratified random sample of webpages for manual testing based on the universe of webpages that had previously gone through automated testing. Embedded algorithms calculate standard errors and the resulting lower- and upper-bound compliance rate for domains.

This paper makes the case that this approach is a more efficient way to guide webmasters and compliance managers on where to focus their remediation efforts.

Introduction

Automated testing can only assess up to 30-35 percent of the WCAG standards and as a result, require manual accessibility testing. The challenge is testing all of a web product's content would be cost prohibitive and time consuming. Randomly selecting a statistically representative sample of webpages for manual testing can provide insight into web products' compliance rates but it does not provide insight about how to prioritize remediation efforts or the risks associated with non-compliance.

Optimal has designed a next generation approach by randomly selecting a stratified sample of web content for manual testing based on the following risk factors or strata:

- automated WCAG standard testing results
- volume of user traffic
- structure or format of webpages[1]
- frequency and intensity of web page structure/format/content changes

Embedded algorithms, designed by statisticians and data scientists, would enable iAccessible to generate risk-based strata, based on the outlined risk factors, from the population of an organization's webpages. The webpages within the sampled strata could then be manually tested and greater insights into high areas of non-compliance and where to allocate resources for remediation efforts could be provided to webmasters and compliance managers.

The remainder of the paper is structured as follows:

- Literature review
- Random sampling approach
- Proposed methodology
- Conclusions and next steps

Literature Review

A similar paper by Rowland and Joeckel, titled "Validating a Sampling Process for Automated Accessibility Testing of Websites in a National Network", from the 2018 ICT Accessibility Symposium, looked at using automated testing and a simple random sampling technique to answer a dual research question. First, how do errors detected in a small sample of web pages correlate with the errors detected on the entire website? Second, how many webpages are needed to construct a 95 percent confidence interval of the entire website?

To answer their dual research question, the authors created two data sets based on 60 websites they deemed to be appropriate. For the first data set, they created a small sample of 14 webpages, for each website. Their second data set consisted of all the webpages for the 60 websites.

For the first portion of the research question, the authors correlated the errors in the small sample to the large sample. The average correlation, for all 60 websites, was 0.57. However, when they grouped the websites based on number of pages, they found the correlation fell as the number of webpages on a website increased.

To answer the second research question, the authors used a bootstrapping method. After calculating the means, they found that 42 percent of all websites' webpages had to be scanned to create the desired 95 percent confidence interval.

The authors mention the added benefit of manual accessibility testing but lamented that the labor intensity of such a venture would make it cost prohibitive. Optimal believes that its methodology

[1] Webpage format may include the following categories: 1. Zig-Zag, 2. F layout, 3. Full Screen Photo, 4. Grid Layout, 5. One-Column Layout, 6. Featured Image Layout, 7. Asymmetrical Layout, 8. Split Screen Layout, 9. Headline and Thumbnails Gallery Layout, 10. Modular Layout, 11. Single Page Layout, 12. Radial Symmetry Layout

of automatically selecting a sample of webpages for manual accessibility testing combined with an integrated manual testing workflow and reporting functionality could address this problem.

Random Sampling Approach

The first portion of the dual research question in Rowland and Joeckel's paper required the creation of a sample. They created the sample by identifying the home page for each of the 60 websites and then using a random number generator to select an additional 13 pages for a total sample of 14 webpages for each of the 60 websites. The total sample size was therefore 840 webpages (60 multiplied by 14) and since the mean average per website, in their population, was 644 their sample constituted 2.17% of the entire population (844 divided by 38,640) (Rowland and Joeckel, 2018).

The above methodology is an example of a random sampling approach, specifically simple random sampling without replacement (SRSWOR) (Pathak, 1988).

For SRSWOR, observations are selected from a population in such a way that every observation in the population has the same probability of being selected for the sample. Sampling is stopped when the total sample size is reached, and no observation can be picked twice for the sample. This is the case in the above example for the 13 webpages randomly selected by Rowland and Joeckel.

The figure below depicts a visual representation of this type of sampling (Thompson,2012):

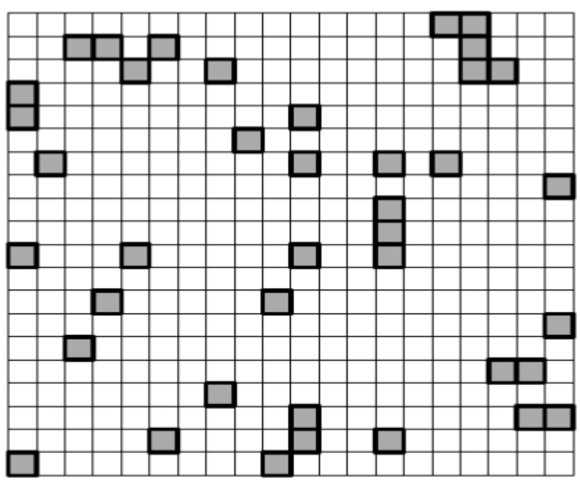

Figure 1: Random sample of 40 observations from population of 400

SRSWOR sampling is beneficial in that it is conceptually simple and, when done correctly, it can result in unbiased estimates of population parameters. It, however, does not provide insight about how to prioritize remediation efforts or the risks associated with non-compliance. This weakness is particularly evident if the underlying data is skewed.

In the specific scenario of sampling webpages for accessibility issues, skewness of data are a reasonable assumption. For example, it is reasonable to assume that webpages that are dynamic, i.e. constantly changing content, would have higher rates of WCAG testing errors than webpages that are static.

To illustrate why skewness of data are important please see the figure below:

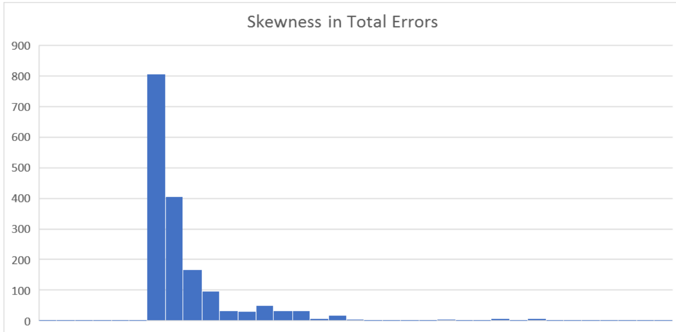

Figure 2: Total 508 Errors on Website

The histogram is produced using data from an Optimal client. Any identifiable information was removed in order to maintain confidentiality. Using WAVE, Optimal conducted automated WCAG tests of 1,727 webpages of a client website. The histogram shows the distribution of "total errors" (errors plus contrast errors as identified by WAVE).

We see visually from the histogram that the data are skewed. Calculating the skewness measure from the population, we get 3.68 which indicates highly skewed.

In deriving a sample of webpages for manual testing using this approach we would only gain insights into the compliance rates of these webpages as a whole but not much else about important components or strata within the population. For example, in this population, there are 34 webpages with greater than 150 total errors based on automated WCAG testing. Since in random sampling each observation has the same probability of being chosen, assuming a sample size of 100, we would only expect two of these high error pages to be included in the sample. This is problematic within the context of accessibility testing. In this example, the highest number of automated WCAG errors on a webpage is 237; a simple random sampling approach does not differentiate selection between webpages that had zero detectable errors and those that had many (in this case 237) errors as determined by automated WCAG tests.

To reiterate, using a simple random sampling approach is appropriate when clients are only interested in the overall rates of compliance. However, if clients are interested in insights on how to focus remediation efforts then using a stratified sampling approach may be more cost-effective despite being more difficult to administrator than simple random sampling. The methodology for stratified random sampling is described in the proceeding section.

Proposed Methodology

This paper outlines two distinct approaches determining strata or assignment to a subgroup and then the sample size determination for each stratum.

Proportional stratification. In this case, the size of the strata is proportionate to the size of the subgroup in the population. For example, if strata were based on webpage structure and it was determined that a domain used had three distinct layouts (e.g., Featured Image Layout, One-Column Layout, Modular Layout). During the automated testing phase, embedded algorithms determine distinct types of webpage layouts or formats and the proportion within the domain. Based on these parameters and the selected level of confidence (say 95%) the resulting sample size is determined.

Webpage Layout Stratum	Population size	Standard deviation	Mean
Featured Image Layout			
One-Column Layout			
Modular Layout			

Disproportionate stratification. With disproportionate stratification, the sample size of each stratum does not have to be proportionate to the population size of the stratum.

iAccessible Algorithms and Workflow

Implementing these sampling approaches outlined above are technically complex. Embedded algorithms and configured workflows in the iAccessible platform attempt to simplify the implementation of risk-based stratified random sampling approach that includes the following steps:

1. *Universe of webpages*. A webcrawler scans and catalogs webpages within the bounds of the targeted domain to a specified depth level.
2. *Construct analytic data table*. The cataloged list of webpages is joined with other data that include risk-based metrics including:
   - automated WCAG standard testing results;
   - volume of user traffic;
   - assessment and categorization of web page structures; and,
   - frequency and intensity of web page structure/format/content changes.
3. *Select risk factors*. User selects the specific subset of risk factors.
4. *Pre-defined strata*. iAccessible's embedded algorithms create pre-defined strata based on distribution levels (e.g., 90^{th} percentile of user volume).
5. *Manual testing assignment*. iAccessible automatically assigns selected webpages to the reviewer pool.
6. *Reporting*. iAccessible automatically generates compliance reports that includes confidence intervals for both automated and manual testing results.

Conclusions and Next Steps

Stratified random sampling has the potential to generate more precise manual testing results than manual testing that uses simple random sampling. Moreover, using stratified random sampling also enables clients to identify specific subgroups are resulting in violating WCAG standards.

The following conditions must be met to select a sampling based on stratified random sample:

Automated scrawl and assessment or other automated assessments much be able to identify each subgroup or strata and classify each of them into only one subgroup. Defining an exhaustive and definitive list of options and categorizes them in an automated fashion in some cases may be difficult (e.g., webpage layouts).

Implementing an approach that involves two or more strata simultaneously becomes increasing complex and increases the resulting sample size to achieve the same level of confidence.

Next Steps. Optimal is currently piloting single strata sampling approach on the Revelo-powered iAccessible platform.