An Exploration of Sample Sizes for Content Analysis of the New York Times Web Site

By Xiaopeng Wang and Daniel Riffe

Introduction|Literature Review|Method|Results|Conclusion and Discussion

Abstract

This study explores the effectiveness and efficiency of sample sizes for content analysis of online newspaper sites. Using simple random sampling, the comparisons showed that a sample size of six days was effective and efficient to represent one year of content of the New York Times Online. Generalizing theresults to content analyses of other online news sites may require special caution due to Web sites' varying formats and contents.

Introduction

Internet communication researchers face a number of methodological challenges.[1]Compared to traditional media research, where quantitative methods prevailed for many years[2], early Internet research used non-quantitative methods more frequently than quantitative methods.[3] Some argued that difficulty in collecting online data was one of the main obstacles, particularly for quantitative content analyses.[4] The Web's hypertextual and interactive content complicates sampling, unitization, generalizability, and coding processes.[5] Many of those methodological problems need to be addressed. In particular, McMillan urged researchers to investigate the validity of multiple sampling methods on the Web.[6]

The goal of sampling is to generate a manageable subset of data from a large population or a sampling frame to represent that population. An ideal sample involves a tradeoff between the ease of study and the representativeness of the population.[7] An effective sample should represent the population and yield certain results, while an efficient sample uses the smallest data set to achieve the same results.[8] Thus, content analysts should determine how to define a tangible sampling frame, how to draw

a representative sample from the sampling frame, and how large the sample size must be to be not only effective but also efficient.9

On the Internet, the enormous amount of information continuously expands at an exponential rate, and the decentralized nature of cyberspace allows any Web user to create and transmit all forms of information anytime from anywhere. As a result, estimating the sampling frames for Web content analyses is a challenge.10 Riffe, Lacy and Fico likened the Internet to "a city without a telephone book or map to guide people."11

Sample size is another major consideration content analysts have to deal with when sampling the Web. Only a few studies have tested sample sizes for Web content analyses. Hester and Dougall compared different sampling methods and sample sizes on Yahoo!News and suggested using a minimum of two constructed weeks to represent a population of six months of online news content; variables with high levels of variability needed at least five constructed weeks for six months of content.12

The various forms of news sites and the complexity of online news content require further investigations on sample sizes for the Web. Online news sites that are affiliated with traditional media outlets, such as the much-visited New York Times Online, retain a large audience.13 Little research has explored the sample sizes for those online news sites. Therefore, this study will compare the effectiveness and efficiency of sample sizes for content analyses of online news sites, specifically the New York Times Online.

Literature Review

A great number of Internet-related studies have analyzed online news sites, including newspaper sites, television sites, and Web-only news sites. For instance, Li studied newspapers' Web page design;14 Cassidy compared Web-only news sites and daily paper sites;15 Schwalbe examined U.S. news sites' portrayal of the Iraq war.16 Unlike personal blogs or news aggregation sites (e.g. Yahoo!News and Google news), online news sites usually feature an independent newsgathering and editing system, update on a daily basis, and retain a large audience.17 Content analyses of online news sites concentrate mostly on articles, images, or Web pages within specific Web sites. Thus researchers are able to estimate the sampling frames—the entire set of articles or Web pages within a certain time frame. With a definable sampling frame, determining an effective and efficient sample size becomes feasible and worthwhile.

Exploring Sample Sizes

Obtaining an ideal sample size is, in effect, a "cost-benefit question."18 An effective and efficient sample size is achieved at the point when increasing the number of cases will not significantly improve the representativeness of the sample results but decreasing the number will significantly damage the sample's validity. This basic principle underlies most of the work that has been done on sample sizes and methods in content analysis.19 An effective and efficient sample size in content analysis research helps the researcher avoid the cost of analyzing a vast amount of data, a virtue especially relevant in the case of the overwhelming amount of online data, and simultaneously reduces sampling error and helps ensure reasonable validity of the inferences and predictions.20

Half a century ago, research was conducted to disclose the effective and efficient sample size for analyzing newspaper content. Stempel compared samples of 6, 12, 18, 24, and 48 issues of a daily newspaper and discovered that 12 issues from two constructed weeks could effectively represent the content of an entire year.21 Riffe and colleagues later compared simple random sampling, stratified or constructed-week sampling, and consecutive-day sampling of a local daily and also found that two constructed weeks could adequately and effectively represent the population, and that daily-stratified sampling was far more efficient than simple random sampling.22 In addition to examining daily newspapers, Lacy, Robinson, and Riffe investigated sampling of weeklies as well; they found that a random selection of 14 issues from a year or one issue from each month in a year (a stratified sample) would efficiently predict a whole year.23 To assess the sampling of multi-year newspaper studies, Lacy, Riffe, and colleagues continued to explore the best sampling strategy for studying five years of dailies, and their conclusion was that a nine-constructed-week sample, rather than a ten-constructed-week one, was adequate to provide a valid inference to the content of a daily newspaper during five years.24

Sampling studies were not limited to newspaper research. Riffe, Lacy, and Drager used Newsweek and found that a random issue selected from each month was the most efficient method to sample one year's news magazine content.25 Because there are a number of television content analyses, researchers attempted to figure out the most effective method for sampling network news as well. Riffe et al. sampled ABC and CBS newscasts to compare simple random, monthly stratified, and quarterly/weekly stratified sampling over a year.26 They found that the most effective approach was to randomly draw two days per month for a content analysis of one year's broadcast news.27

In virtually all of these studies, the underlying goal was to determine if there are systematic or even "cyclic" factors that can affect sample representativeness, and that which therefore must be controlled or used as "stratifying" variables (e.g., stratifying by day of week, and sampling proportionally among the Mondays, among the Tuesdays, among the Wednesdays, etc.).

At present, researchers' knowledge about content or content-posting cycles on the Web, and their effect on sample representativeness, is limited. No sampling guidelines exist for researchers to select an adequately effective sample in examining content on the Web. Instead, content analysts have applied various methods in their longitudinal research. Some have adopted traditional media sampling strategies. For example, Pitts used a one-constructed-week sample to examine television Web sites over

a six-month period of time;28 Craft and Wanta drew a constructed-week sample as well to represent one month of articles on news sites;29 and Lim used two-constructed weeks between January 1 and December 31 of 2003 to study three news sites' content. However, some of these sampling decisions seem somewhat arbitrary in the absence of empirical testing.30

For instance, Li argued that the Web page designs of news sites were relatively stable, and therefore he studied three U.S. newspaper sites on ten continuous days;31 Pashupati and Lee randomly selected six days in April and May to compare online advertising in Indian and Korean newspaper sites;32 and Schwalbe analyzed Web sites' snapshots every Wednesday during the first five weeks of the Iraq war.33 Boczkowski and de Santos chose two days per week during a ten-week period to represent three online newspapers' coverage between September and November of 2005.34

Although the sampling methods chosen varied depending on specific research questions in these studies, such variation suggests that content analysts might benefit from guidelines for sampling the Web or at least some guidance in determining an appropriate sample size in analyzing Web content.

Research Question

New and rapidly renovated features of the Internet make content analysis of the Web extremely complicated. Even though a longitudinal analysis of selected news Web sites may not involve the problem of defining sampling frames, the effectiveness and efficiency of sample sizes remain crucial concerns. Previous studies on news sites applied various sampling sizes, but lacked justification for claims of the sample's representativeness. As a result, the main research question of this study is simply to find a sample size that is not only effective but also efficient to examine the New York Times online.

Method

Prior sampling explorations commonly required three steps: (1) creating the population parameters on several variables; (2) drawing different samples of different sizes using different sampling strategies; and (3) comparing sample statistics on the variables with population variable parameters to determine which samples are most effective.35 This study followed this procedure.*Creating a Parameter*

To create a population parameter, the online version of the New York Times, NYTimes.com, was chosen for this study. The New York Times is frequently referred as the elite U.S. paper, "newspaper of record," and the "agenda setter" for other print and electronic media.36 With 21.5 million unique visitors per month, NYTimes.com is the leading newspaper site in America.37 In addition, NYTimes.com appears to consistently update its pages throughout the day, evidenced by inclusion of the time of

update next to each headline, which suggests that the NYTimes.com is an ideal focus for a sampling study of this nature. As in many previous Web content analyses,38 this study examined the snapshots of NYTimes.com's front page, especially focusing on its top headline portion. A Web site's front page serves as the front door to the site, like the newspaper's front page,39 and the top headline portion, the center part of a Web site's front page excluding side menu and advertisements, is the most prominent position on a site, offering users the first impression of a site.40 Under a normal display setting with a resolution higher than 800 by 600 pixels, the top headline portion could be viewed on the first screen in a computer monitor.

The top headline portion usually presents the most newsworthy headlines and photographs, is most frequently updated, and thus has gained particular attention from many Internet content analysts.41 Moreover, the HTML codes of the top headline portion were more flexible than in the rest of the page, so Web editors were able to adjust content with minimum technique constraints. Measuring the variations of content in the top headline portion would well reflect an online newspaper's editorial decision making.

Variables

Variables for this study included the story topics, geographic bias, number of links, and uses of multimedia in story presentation. All variables are common to content analyses of Web sites and were measured within the top headline portion of the front page of NYTimes.com.

The categories for the story topic variable were adopted from Stempel's frequently-cited study, which sorted news stories into political and government acts, war and defense, diplomacy and foreign relations, economic activity, agriculture, transportation and travel, crime, public moral problems, accidents and disasters, science and invention, public health and welfare, education and classic arts, popular amusements and general human interest.42 Eventually, the researcher would use this information to calculate a daily percentage of war news to compare multiple sampling sizes in terms of that particular percentage.

The geographic bias variable, measuring the unevenness of coverage of nations and geographic areas in the news, was adapted from Mayo and Pasadeos' classification of the world: the United States, the U.S. neighbors, Central/South America, Western Europe, Eastern Europe, Mid-East and North Africa, Africa (Sub-Sahara), South Asia, Japan, the Four Tigers, other East Asia, and Oceania.43However, the "Four Tigers" designation (which referred to Singapore, Malaysia, Taiwan, and South Korea) was eventually dropped, because it was no longer considered a unique geographic region in the 21st century. For comparison of sample to population parameter, the researchers examined the number of U.S. domestic stories divided by the other categories to yield the percentage of foreign news.

Hyperlinks and multimedia are two new types of content that the Internet employs.44 The hyperlinks variable included every internal or external link that pointed to another destination page or file. And multimedia referred to non-text format of information, including images, video, audio, and interactive features.45The variable uses of multimedia was calculated as the ratio of the number of multimedia and the number of hyperlinks.

Estimating parameters

The timeframe for this study was one year, from July 6, 2005, to July 5, 2006. To estimate the parameters for the year, the researcher logged on to the NYTimes.com and captured snapshots of the front page on a daily basis. Previous Web content analyses normally coded one snapshot per day to represent the Web content in a 24-hour cycle. Given that the Web was updated continuously, instantly, and therefore irregularly, a pilot study was conducted to explore how many snapshots per day would be effective and efficient to estimate the parameter in 24 hours. In the pilot study, assuming one snapshot would be extensive and sufficient to detect Web content variability every hour, the researcher captured a snapshot of NYTimes.com's front page every hour within a consecutive week, from July 24 to 30, 2005. Accordingly, a total of 168 snapshots were obtained to calculate a parameter for one week. Afterward, the researcher compared 50 sets of simple random samples of seven and 14 snapshots a week, as well as 50 sets of constructed-week samples and found that both the simple random sample of seven snapshots a week and one stratified-week sample could effectively predict the one-week population. In other words, one snapshot per day of NYTimes.com's front page could be sufficient to represent each day's Web content.

For consistency purposes, the researcher logged on to the NYTimes.com every noon and evening, around 11 p.m., captured a snapshot of the front page, and saved it to the computer's hard drive. The mean of the two observations per day was calculated to represent each day, and was then used by the researcher to calculate the population parameter for the one-year NYTime.com front pages. Due to the network accessibility and other reasons, 23 days' snapshots were not successfully saved during the research time frame. This study eventually collected 684 snapshots of 342 days within a year.

Intercoder reliability

A student and the first author coded 98 snapshots, 14% of the entire sample, for the intercoder reliability test. According to Holsti's formula,46 the simple agreement was 80.0% for Story Topic, 95.1% for Geographic Bias, 98.2% for Number of Hyperlinks, and 86.7% for Uses of Multimedia. The researcher coded the remaining 586 snapshots of the one-year NYTimes.com front pages.

Multiple sampling

To compare sampling sizes, the researchers separated the 684 snapshots into two individual observations—noon and evening—based on the time when the snapshots were taken. By using noon and evening observations, the researchers, in effect, manipulated two sets of data and compared sampling methods twice, which was expected to enhance the reliability of the comparison results.

Simple random samples of size three, four, five and six days of snapshots were drawn from the noon and the evening observation of the site. Fifty samples were drawn for each sample size and each noon and evening observation. Therefore, a total of four sets of 50 samples were chosen for each noon and evening observation.

Sample Size Comparison

The sample means and sample standard errors were calculated for five variables for all the samples. Each sample size was tested to see whether the population means for five variables fell into one and two standard errors from the sample means. The Central Limits Theorem predicts that the sample mean distribution is close to a normal curve and the mean of the samples means is approximately the population mean. Thus, in 95% of samples, the population mean should be between two standard errors of the sample mean, and in 68% of samples, the population mean should fall within one standard error of the sample mean.47 Accordingly, a sample size is considered effective only if its percentage exceeds or equals these critical percentages; a sample size is efficient only if the next smaller sample size does not meet the percentage standards.

Results

The population

As shown in Table 1, during the one-year period of time, from July 6, 2005, to July 5, 2006, NYTtimes.com published approximately 5.17 (SD = 1.38) U.S.-related stories every day on the top headline portion of its front page. In contrast, merely 34.29% (SD = 14.67) of the top headlines covered the rest of the world. Of the everyday top headlines, the coverage of war, defense, and terrorism accounted for 11.79% (SD = 10.62). On average, there were 20.45 hyperlinks (SD = 4.97) pointing to other Web pages or files, among which 15.90% (SD = 7.802) were in multimedia formats other than text-based content.

Table 1

Population Distribution for NYTimes.com Front Page Coverage of U.S., Foreign, and War News, Number of Hyperlinks and Percentage of Multimedia, July 6, 2005 to July 5, 2006.

	Mean	*SD*	Coefficient of Variation
Number of U.S. news	5.17	1.38	.266
Percentage of foreign news	34.29	14.67	.428
Percentage of war news	11.79	10.62	.901
Number of hyperlinks	20.45	4.97	.243
Percentage of multimedia	15.90	7.80	.490

The coefficient of variation is the standard deviation divided by the mean, indicating the variability of units in a population. The higher the coefficient, the more variable the population is. Sampling researchers usually examine coefficient of variation to test the variability assumption and detect the impact of variability on sampling size.48 In particular, if the coefficient of variability exceeds .5, researchers advise increasing the size of the sample.49 The coefficient of variation for the percentage of war news was the highest (.901) compared to the other four variables, according to Table 1.

The samples

To answer the research question about finding an effective and efficient sample size, comparisons of multiple sampling predictions with the population parameter were conducted as shown in Table 2. According to the Central Limits Theorem, 95% of random sample means will be within plus or minus two standard errors of the population mean, and 68% will be within plus or minus one standard error of the population mean.

Comparing the noon and evening observations' results for 50 samples for each of size three, four, five, and six simple random sample days, the random selection of six days was found to be the most effective and efficient sample size. The three-day sample was apparently insufficient because for this sample size 11 measurements failed to generate a sample mean that met the Central Limits Theorem standards. Randomly selecting four days greatly improved the sample efficiency, but out of 50 sets of samples only 92% in the noon observation and 94% in the evening observation for the percentage of war news were within two standard errors; 94% in the noon observation and 90% in the evening observation for the variable of number of hyperlinks were within two standard errors; 94% in the noon observation for the number of U.S. news, and 92% in both noon and evening observations for the percentage of multimedia were within two standard errors, which all violated the Central Limits Theorem's assumptions. Furthermore, drawing five random days seemed fairly effective to estimate the one-year population except for the percentage of war news in the evening observation, the number of hyperlinks in the noon observation, and the use of multimedia variable in the evening observation.

Table 2

The Percentage of Random Sample Means Falling within One and Two Standard Errors of Population Mean in Sets of 50 Samples of NYTimes.com Front Page Regarding the Coverage of U.S., Foreign, and War News, Number of Hyperlinks and Percentage of Multimedia, Two Observations during July 6, 2005 to July 5, 2006.

Sample Size	Six		Five		Four		Three	
	1 SE	2 SE	1 SE	2 SE	1 SE	2 SE	1 SE	2 SE
Number of U.S. news								
Noon observation	98	100	88	98	84	~~94~~	76	~~92~~
Evening observation	88	98	82	96	80	~~94~~	76	~~92~~
Percentage of foreign news								
Noon observation	92	100	90	100	88	98	78	~~92~~
Evening observation	94	98	88	96	86	96	76	~~88~~
Percentage of war news								
Noon observation	96	98	80	96	82	~~92~~	74	~~90~~
Evening observation	88	98	84	~~94~~	80	~~94~~	~~66~~	~~92~~
Number of hyperlinks								
Noon observation	92	98	84	~~94~~	72	~~94~~	70	~~86~~
Evening observation	84	96	82	~~94~~	78	~~90~~	76	~~90~~
Percentage of multimedia								
Noon observation	96	100	84	96	82	~~92~~	86	~~94~~
Evening observation	90	98	84	~~88~~	86	~~92~~	86	~~92~~

Note: Crossed percentages indicate the sample means did not meet 68% or 95% critical values.

The criteria predicted by the Central Limits Theorem were not fully met until the sample size was enlarged to six. Considering the average variations as shown in Table 3, a sample of six days yielded a smaller variance, for every variable, than a sample of five days. For sampling purposes, a smaller variance is more desirable in terms of reducing sample errors. Therefore, a sample size of six days will effectively and efficiently represent the content on the NYTimes.com in a one-year period of time.

Table 3

Average Variances of Number of U.S. News, Percentage of Foreign News, Percentage of War News,

Number of Hyperlinks, and Percentage of Multimedia for Sample Size of Four, Five and Six Random Days of NYTimes.com Front Page

Sample Size	Six	Five	Four	Population
Number of U.S. news				
Noon observation	1.75	1.79	2.56	1.96
Evening observation	1.85	2.03	1.97	1.83
Percentage of foreign news				
Noon observation	208.91	241.79	271.07	234.10
Evening observation	193.20	196.60	197.71	203.61
Percentage of war news				
Noon observation	11.07	12.67	9.54	11.29
Evening observation	8.63	11.08	12.18	11.41
Number of hyperlinks				
Noon observation	25.08	29.09	27.31	26.96
Evening observation	11.91	21.38	24.16	21.93
Percentage of multimedia				
Noon observation	54.29	56.68	67.98	60.53
Evening observation	51.59	55.13	74.72	61.10

Conclusion and Discussion

This study found that a sample size of six days was effective and efficient for an analysis of the New York Times Online content during one year. Generalizing the results to other Web content analyses may need special caution because of the enormous variations in Web content. This study focused on one type of news site, one that involves an independent newsgathering and editing system and is updated regularly.50 The efficient sample size identified here and based on these news sites may not be applicable to other forms of Web sites, such as Web blogs and personal sites. In addition to news Web sites like NYTimes.com, future studies should explore multiple sample size comparisons on more news sites of the same type, and other types. For other types of Web sites, researchers may need more creative approaches to test the sampling size effectiveness and efficiency.

Additionally, although the sample size of six days is not only effective but also efficient for predicting this newspaper site's content of a year, sampling methods may vary depending on different research questions and designs. In the comparisons of sample means and population parameters for five variables, the percentage of foreign news met the Central Limits Theorem's criteria in all three sample sizes except for the three-day sample, and the percentages of sample means of the number of U.S. stories met the criteria in six-day and five-day samples. However, some variables like the use of multimedia seemed to be much more "sensitive" than others in the sampling tests because among 50

sets of the five-day samples only 88% sample means fall within the two standard errors of the population mean. Such a contrast reveals the importance of taking practical research questions into account when determining a sample size for content analysis. It is risky to choose a sample size based on a guideline if one is unaware of the "sensitiveness" or variability of the variables.

However, research questions and research designs are diverse in many ways. For instance, previous sampling research suggests increasing sample size if the coefficient of variation is greater than .5.51 In this study, however, more trouble was caused by variables with relatively low variability rather than those with high variability. A similar pattern was found in previous studies as well. Lacy et al. identified one set of samples that failed to meet the criteria as the one with the least amount of variation in their examination of seven constructed-week samples for multi-year studies.52

These inconsistent findings suggest that researchers might need to examine other assumptions in addition to variability. The Central Limits Theorem predicts that things in nature tend to be normally distributed, including the distribution of means. Regardless of the normality of the population distribution, the shape of the sample means of the population is approximately normal.53 Vogt also pointed out that the distribution of sample means would be much closer to a normal curve if the sample size were 30 or more.54

With a larger sample size, the feature of a normal distribution would be more apparent. In this sampling study, sample sizes for comparison were three, four, five, and six days, obviously much smaller than 30. Even though the small sample size did not preclude use of the Central Limits Theorem, it could be an explanation of the inconsistent relations of variability versus sample sizes in sampling studies.

As an exploration of sample sizes, this study only compared different simple random samples for analyzing newspaper sites. The result of using six simple random days to represent one year could not only serve as a guideline for future Web content analysis, but also serve to initiate further investigations on different types of sampling methods, such as the stratified sampling that some Internet studies used. Prior sample size research applied stratification to sampling newspapers and magazines to reduce variance because their amount of news varies considerably by cycle. This study could use simple random samples because the amount of news in the sample does not differ too much from day to day. The average number of stories on the front page's headline portion was 20.45 (SD=4.97, CV=.243). When the variance increases, stratified sampling will be needed. As shown in Table 3, a simple random sample of six days does not reduce the variances significantly compared to the population. Simple random samples of 12 and 24 days were further tested to compare the average variances for all the variables. However, the variation was not reduced considerably. Future sample size comparisons can employ other sampling techniques. Given the routines in traditional journalism,55 a weekly news cycle is expected to occur in online newspaper sites. In addition to comparing simple random samples, future research may test sample sizes' efficacy and efficiency when sample the Web using constructed weeks, as well as other sampling methods.

Due to limitations of data, this study could only create a population parameter for one-year period of time. Future studies must extend the time frame to multiple years, when a larger sample size may be required, as Lacy et al. showed.56 Then, more complex sampling comparisons, including the contrasting of different types of sampling strategies, could be conducted. On the other hand, with a large population, the comparison results and the recommendations for real world content analyses will be more valuable and practical.

Xiaopeng Wang is a faculty member in the Department of Journalism and Media Studies at the University of South Florida St. Petersburg. Daniel Riffe is the Richard Cole Eminent Professor in the School of Journalism and Mass Communication at the University of North Carolina at Chapel Hill