

EXPLORING SAMPLE SIZES FOR CONTENT ANALYSIS OF
ONLINE NEWS SITES

Xiaopeng Wang

Submission to the Communication Theory & Methodology Division, AEJMC
August, 2006

EXPLORING SAMPLE SIZES FOR CONTENT ANALYSIS OF ONLINE NEWS SITES

Introduction

Studies of research trends predict that the Internet will continue to be an important topic in communication scholarship.¹ In the past two decades, Internet-related research has been moving from early descriptive studies about the medium itself--the Internet's characteristics--to a higher level, focusing on Web users and social effects.² Kopper and colleagues, for instance, have identified seven perspectives that current Internet research has been pursuing, including analyses of product, users, quality, social context, market, occupational changes, and experimental projects.³

Meanwhile, Stempel and Stewart also point out methodological challenges that communication researchers had to deal with, especially in audience research and content analysis.⁴ In contrast to the traditional media studies, where quantitative research has prevailed for many years,⁵ Kim and Weaver discover that early Internet research used non-quantitative methods more frequently than quantitative methods.⁶ Some assumed that

¹ Rasha Kamhawi and David Weaver, "Mass Communication Research Trends: From 1980 to 1999," *Journalism & Mass Communication Quarterly* 80(Spring 2003): 7-27.

² Sung Tae Kim and David Weaver, "Communication Research about the Internet: A Thematic Meta-analysis," *New Media & Society* 14(4) 2002: 518-538.

³ Gerd G. Kopper, Albrecht Kolthoff, and Andrea Czepek, "Research Review: Online Journalism—A Report on Current and Continuing Research and Major Questions in the International Discussion," *Journalism Studies* 1(Number 2000): 499-512.

⁴ Guido H. Stempel III and Robert K. Stewart, "The Internet Provides both Opportunities and Challenges for Mass Communication researchers," *Journalism & Mass Communication Quarterly* 77(Autumn 2000): 541-548.

⁵ Kamhawi and Weaver, "Mass Communication Research Trends."

⁶ Kim and Weaver, "Communication Research about the Internet."

difficulty in collecting online data was one of the main constraints preventing quantitative studies;⁷ in particular, they point to sampling problems in content analysis.⁸

While efforts have been made to explore to take advantage of the Internet to conduct online surveys,⁹ content analysis continues to face many challenges in measuring the hypertextual and interactive Web content,¹⁰ as well as problems of sampling, unitization, and coding. The purpose of this study is to explore one of the problems, comparing the effectiveness and efficiency of different sample sizes for content analysis of the World Wide Web.

Literature Review

Sampling is a problem

Researchers have discovered a number of methodological problems involved in content analysis of the Internet. Gunter argued that the nonlinearity and customized settings of Web content violated the underlying assumptions of the traditional content analysis method because Web users could read across stories and even Web sites through the hyperlinks.¹¹ Neuendorf noted that the Internet combined the existing media's

⁷ Kamhawi and Weaver, "Mass Communication Research Trends."

⁸ Stempel and Stewart, "The Internet Provides both Opportunities and Challenges."

⁹ Samuel J. Best and Brian S. Krueger, *Internet Data Collection*, Thousand Oaks, CA: Sage, 2004; Matthias Schonlau, Ronald D. Fricker, and Jr., Marc N. Elliott, *Conducting Research Surveys via E-mail and the Web*, Rand Publications, 2001. Available online at <<http://www.rand.org/publications/MR/MR1480/>>.

¹⁰ Barrie Gunter, *News and the Net* (Mahwah, NJ: Lawrence Erlbaum, 2003); Sally J. McMillan, "The Challenge of Applying Content Analysis to the World Wide Web," *Journalism & Mass Communication Quarterly* 77(Spring 2000): 80-98; Stempel and Stewart, "The Internet Provides both Opportunities and Challenges;" Christopher Weare and Wan-Ying Lin, "Content Analysis of the World Wide Web: Opportunities and Challenges," *Social Science Computer Review* 18(Fall 2000): 272-292.

¹¹ Gunter, *News and the Net*.

features with hypertextuality, interactivity, and many other critical attributes. The mix of various media characteristics made Internet content analysis fairly complex.¹² Stempel and Stewart were concerned with how such complexity affected generalizability and representativeness regarding the Web content analysis.¹³

McMillan concluded that five major issues exist when examining Web content:

- 1, how to identify the units to be sampled?
- 2, how to collect data for cross-coder tests when the Web changes rapidly?
- 3, how to solve copyright issues if researchers download Web pages for analysis?
- 4, how to standardize units of analysis given the multimedia features of the Web?
- 5, how to check inter-coder reliability?¹⁴

Similarly, Weare and Lin examined the potential methodological issues of content analysis and identified problems existing in the processes of sampling, unitization, categorization, and coding.¹⁵ In particular, McMillan recommended that researchers investigate the validity of multiple sampling methods on the Web.¹⁶

The goal of sampling is to generate a manageable subset of data from a large population or a sampling frame to represent this population. An ideal sample is a tradeoff between the ease for study and the representativeness of the population. 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.¹⁷

¹² Kimberly A. Neuendorf, *The Content Analysis Guidebook* (Thousand Oak CA: Sage, 2002).

¹³ Stempel and Stewart, "The Internet Provides both Opportunities and Challenges."

¹⁴ Sally J. McMillan, "The Challenge of Applying Content Analysis to the World Wide Web."

¹⁵ Weare and Lin, "Content Analysis of the World Wide Web."

¹⁶ McMillan, "The Challenge of Applying Content Analysis to the World Wide Web."

¹⁷ Klaus Krippendorff, *Content Analysis* (Beverly Hill, CA: Sage, 1980).

Dominick complained that there was no existing sampling frame on the Web.¹⁸ On one hand, the amount of information on the Web is enormous and expands at an exponential rate. On the other hand, the decentralized nature of cyberspace allows any Web user to create and transmit various forms of information anytime from anywhere. The anonymity makes it even harder to estimate the sampling frames for content analytic research.¹⁹ No wonder Riffe, Lacy and Fico described the Internet as “a city without a telephone book or map to guide people”.²⁰

In practice, content analysts commonly used online search engines and available directories for their sampling frames. For instance, Dominick located 500 personal home pages via Yahoo! Directory;²¹ Paul found 64 disaster relief home pages by using several online search engines;²² and Liu et al. analyzed business Web sites by using the *Fortune* 500 companies index.²³ However, using research engines and assorted directories was still problematic because Web sites emerge and recede too rapidly to be traced, choosing key words for searching is a tricky business, and even the most sophisticated search

¹⁸ Joseph R. Dominick, “Who Do You Think You are? Personal Home Pages and Self-Presentation on the World Wide Web,” *Journalism & Mass Communication Quarterly* 76(Winter 1999): 646-658.

¹⁹ Weare and Lin, “Content Analysis of the World Wide Web.”

²⁰ Daniel Riffe, Stephen Lacy, and Frederick G. Fico, *Analyzing Media Messages: Using Quantitative Content Analysis in Research* (Mahwah NJ: Lawrence Erlbaum, 1998), p.101.

²¹ Dominick, “Who Do You Think You are?”

²² Mary Jae Paul, “Interactive Disaster Communication on the Internet: A Content Analysis of Sixty-Four Disaster Relief Home Pages,” *Journalism & Mass Communication Quarterly* 78(Winter 2001): 739-753.

²³ Chang Liu, Kirk P. Arnett, Louis M. Capella, and Robert C. Beatty, “Web Sites of the Fortune 500 Companies: Facing Customers through Home Pages,” *Information & Management* 31(1997): 335-345.

engines can find only a small amount of information online. Some claimed the size of the “invisible” data to search engines was 500 times the content that could be searched.²⁴

Sample size is another issue involved in sampling of the Internet, which, however, has drawn little attention from Internet content analysts. Part of reason might be the difficulty of testing sample size effectiveness because researchers are not even aware of the size of sampling frames, including both visible and invisible data. Thus, the invisibility of Web content toward search engines essentially leads to the fact that researcher can merely draw a convenient sample using search engines. The representiveness of a sample from visible data becomes considerably skeptical.

Questions about sampling method on the Internet might leave researchers feeling hopeless. But before being pessimistic, researchers need to notice not all sampling frames are indefinable on Web, depending on units of analysis. For instance, analyses of a group of Web sites may initially be plagued by troubles in defining the group, because those Web sites are essentially independent or isolated in the cyber world. Researchers cannot detect them unless they are linked together. Search engines have established a form of linkage to gather these Web sites, partially not comprehensively, which causes the problem of defining sampling frames.

On the contrary, longitudinal research designs have an relatively explicit sampling frame, such as Li’s study of newspapers’ Web pages design,²⁵ Massey and Chang’s

²⁴ Javed Mostafa, “Seeking Better Web Searches,” *Scientific American* 292(February 2005):66-68.

²⁵ Xigen Li, “Web Page Design and Graphic use of Three U.S. Newspapers,” *Journalism & Mass Communication Quarterly* 75(Summer 1998): 353-365.

analysis of Asian Web newspapers,²⁶ and Cassidy's comparison of Web-only news sites and daily paper sites.²⁷ Their units of analysis are articles or Web pages within a specific site rather than isolated Web sites. Since articles and Web pages are connected to a Web site, they are all visible. Thus researchers should be able to estimate the sampling frames—the overall articles or Web pages within a certain time frame. Then, whether a sample size is representative should have become researchers' concern, which unfortunately has not been addressed clearly or convincingly.

Consequently, this study concentrates on comparing the effectiveness of different sample sizes in the longitudinal content analysis of daily updated news sites. An overview of the literature discovered that a number of Internet studies of online media outlets, including newspaper sites, television sites, and Web-only news sites. As a major type of online journalism, these online media outlets own an independent newsgathering and editing system, update on a daily basis, and retain a large audience.²⁸ Owing to the longitudinal research design on a specific Web site, it is reasonable to estimate the population/sample frame to compare multiple sampling methods.

Exploring Sample Sizes

The question of an ideal sample size, in effect, is a “cost-benefit question”.²⁹ An effective and efficient sample size is achieved at a point where increasing the number of cases will not significantly improve the representativeness of the sample results, while

²⁶ Brian L. Massey, and Li-jing Arthur Chang, “Locating Asian Values in Asian Journalism: A Content Analysis of Web Newspapers,” *Journal of Communication* 52(December 2002): 987.

²⁷ William P. Cassidy, “Web-only Online Sites More Likely to Post Editorial Policies Than Are Daily Paper Sites,” *Newspaper Research Journal* 26 (Winter 2005): 53.

²⁸ Deuze, “Online journalism.”

²⁹ Klaus Krippendorff, *Content Analysis*, p.69.

decreasing the number will significantly damage its validity. An effective and efficient sample size in content analysis research will save the cost of analyzing a vast amount of data, especially the overwhelming online data, and simultaneously reduce sample error to maintain reasonable validity of the prediction.

A half century ago, research was conducted to disclose an effective and efficient sample size for analyzing traditional media. 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.³⁰ Riffe and colleagues 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.³¹ In addition to examining daily newspapers, Lacy, Robinson, and Riffe investigated sampling of weeklies as well. They found that a random selection of fourteen issues from a year or one issue from each month in a year (a stratified sample) would efficiently predict a whole year.³² To assess the sampling of multi-year newspaper studies, Lacy and colleagues continued to experiment with a best sampling strategy for studying five years of dailies. Their conclusion was that a nine-constructed-week sample, rather than a ten-constructed-week

³⁰ Guido H. Stempel III, "Sample Size for Classifying Subject Matter in Dailies: Research in Brief," *Journalism Quarterly* 29(summer 1952): 333-334.

³¹ Daniel Riffe, Charles f. Aust, and Stephen lacy, "The effectiveness of Random, Consecutive Day and Constructed Week Sampling in Newspaper Content Analysis," *Journalism Quarterly* 70 (spring 1993): 133-139.

³² Stephen Lacy, Kay Robinson and Daniel Riffe, "Sample Size in Content Analysis of Weekly Newspapers," *Journalism Quarterly* 75(summer 1995): 336-345.

one, was adequate to provide a valid inference to the content of a daily newspaper during five years.³³

Sampling studies were certainly 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 represent one year's newsmagazine.³⁴ Since there were a great number of television content analyses, researchers attempted to figure out a most effective method to analyzing network news as well. Riffe and colleagues used ABC and CBS newscasts to compare simple random, monthly stratified and quarterly/weekly stratified sampling over a year. 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.³⁵

No sampling guidelines existed for researchers to select an adequately effective sample in examining the content on the Web. Instead, content analysts applied various methods in their longitudinal research. Some adopted traditional media sampling techniques. For instance, Pitts used a one-constructed-week sample to examine television Web sites over a six-month period of time;³⁶ and Craft and Wanta drew a constructed-

³³ Stephen Lacy, Daniel Riffe, Staci Stoddard, Hugh Martin, and Kuang-Kuo Chang, "Sample Size for Newspaper Content Analysis in Multi-Year Studies," *Journalism & Mass Communication Quarterly* 78(winter 2001):836-845.

³⁴ Daniel Riffe, Stephen Lacy, and Michael Drager, "Sample Size in Content Analysis of Weekly News Magazines," *Journalism & Mass Communication Quarterly* 73(autumn 1996):635-644.

³⁵ Daniel Riffe, Stephen Lacy, Jason Nagovan, and Larry Burkum, "The Effectiveness of Simple and Stratified Random Sampling in Broadcast News Content Analysis," *Journalism & Mass Communication Quarterly* 73(spring 1996):159-168.

³⁶ Mary Jackson Pitts, "Television Web Sites and Changes in the Nature of Storytelling," *Simile* 3(August 2003).

week sample as well to represent one month of articles on news sites.³⁷ Some seemingly made their sampling decision arbitrarily. For instance, Li argued that the Web page designs of news sites were relatively stable, and therefore he studied three U.S. newspaper sites in ten continuous days;³⁸ and Pashupati and Lee randomly selected six days in April and May to compare online advertising in Indian and Korean newspaper sites.³⁹

Although the sampling methods varied depending on their specific research questions in these studies, such variation suggest that content analysts needed a sampling guideline for examining the Web or at least some assumptions about sampling need testing with Web content.

Research Question

New and rapidly renovated features of the Internet make content analysis of the Web extremely complicated. Even though an analysis of selected news Web sites may not involve the problem of defining sampling frames, the effectiveness and efficiency of sample sizes may remain a crucial concern. Previous studies about news site content have applied various methods, but lack of an empirical examination of their validity and representativeness. What a sample size is not only effective but also efficient to examine

³⁷ Stephanie Craft and Wayne Wanta, "Women in the Newsroom: Influences of Female Editors and Reporters on the News Agenda," *Journalism & Mass Communication Quarterly* 81(spring 2004): 124.

³⁸ Xigen Li, "Web Page Design and Graphic Use of three U.S. Newspapers."

³⁹ Kartik Pashupati and Jeng Hoon Lee, "Web Banner Ads in Online Newspapers: A Cross-National Comparison of India and Korea," *International Journal of Advertising* 22 (2003): 531.

the content in the online environment? This is the main question that this study attempts to address.

Method

Existing sampling explorations commonly required three steps: (1) creating the population parameters; (2) drawing different samples using different sampling strategies; and (3) comparing which samples were most effective.⁴⁰ This study followed this procedure.

Creating A Parameter

Media To create a population parameter, the online version of the *New York Times*, NYTimes.com, was chosen for this study. As most of previous Internet content analyses, this study examined the snapshots of NYTimes.com's front page. Specifically, the researcher focused on the top headline portion, where is the top of the NYTimes.com's front page, because the Web page design caused the content placement lacked variability in the rest of the front page. For instance, under the top headline portion, NYTimes.com presented three headline stories for each of the following categories: Business, National, Washington, Health, Science, Arts, Movies, Theater, Dining & Wine, Real Estate, International, New York/Region, Sports, Technology, Travel, Books, Fashion & Style, Education, Home & Garden, and Automobiles. It would be insignificant to measure the variance of news topics in this portion.

Actually, many Internet content analysts paid particular attention on top headline portions. Tremayne, for example, coded the "main page stories" appearing on the screen when people initially logged on a Web site, which was similar to the top headline portion

⁴⁰ Lacy, Riffe, Stoddard, Martin, and Chang, "Sample Size for Newspaper Content Analysis in Multi-Year Studies."

in this study.⁴¹ Under a normal display setting with a resolution higher than 800*600, the top headline portion could be viewed on the first screen in a monitor. This portion, functioning as a newspaper front page and a magazine cover, usually presented the most newsworthy headlines and photographs and was most frequently updated. Moreover, the designs and maintenance in this portion were more flexible than in the rest of the page so that Web editors were able to adjust content with minimum technique constraints.

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

The categories of *story topics* 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 calculated the daily *percentage of war coverage* to conduct sampling methods comparison.

The *geographic bias* variable was adapted from Mayo and Pasadeos' widely used classification of the world: the United States, the United States' neighbors, Central/South America, Western Europe, Eastern Europe, Mid-East and North Africa, Africa (Sub-

⁴¹ Mark Tremayne, "The Web of context: Applying network theory to the use of hyperlinks in Journalism on the Web," *Journalism & Mass Communication Quarterly*, 81 (Summer 2004): 237-253.

⁴² Guido H. Stempel III, "Gatekeeping: The mix of topics and the selection of stories," *Journalism Quarterly*, 62 (Fall 1985): 79-96, 815.

Sahara), South Asia, Japan, Four Tigers, other East Asia, and Oceania.⁴³ The researcher determined to drop “Four Tigers”, which referred to Singapore, Malaysia, Taiwan, and South Korea in Asia, because it was no longer considered as a significant geographic region in the twenty-first century. When comparing the sampling methods, the researchers examined *the number of U.S. domestic stories* and collapsed the rest categories for *the percentage of foreign news*.

The *hyperlinks* included every link that pointed to another destination page or file. And *multimedia* referred to non-text format of information, including images, video, audio, and interactive features⁴⁴. The *uses of multimedia* variable was calculated as the ratio between the number of multimedia and the number of hyperlinks.

Estimating a parameter The timeframe for this sampling comparison study was a six-month time period, from July 6, 2005 to January 5, 2006. To estimate the parameter for the six months, the researcher determined to log on to the NYTimes.com periodically and take snapshots of the front page. Previous Internet content analyses normally coded one snapshot per day to represent the Web content in a 24-hour cycle. Given the fact that the Web was updated instantly, continuously, and thus irregularly, a pilot study was conducted to explore how many snapshots per day would be effective and representative to estimate the 24 hours’ parameter.

In the pilot study, assuming one snapshot per hour would be extensive and sufficient to detect Web content variability, the researcher took a snapshot of NYTimes.com’s front page every hour within a consecutive week, from July 24 to 30,

⁴³ Charles Mayo and Yorgo Pasadeos, “Changes in the International Focus of U.S. Business Magazines, 1964-1988,” *Journalism Quarterly* 68(Autumn 1991): 509-514.

⁴⁴ The interactive features found in the NYTimes.com were mainly flash, interactive maps, and readers’ forums.

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 7 and 14 snapshots a week, as well as 50 sets of constructed-week samples and found that both the simple random sample of 7 snapshots a week and one stratified-week sample could effectively predict the one-week population. In other words, taking one snapshot per day of NYTimes.com's front page could be sufficient to represent that day's Web content.

Being conservative and consistent, the researcher logged on to the NYTimes.com every noon and evening, around 22:00 o'clock, took a snapshot of the front page, and saved it to the local desktop. The mean of the two observations per day were calculated to represent each day, based on which the researcher calculated the population parameter for the six-month of NYTime.com's front page. Due to network accessibility and other reasons, 14 days' snapshots were not successfully saved during the research time frame. The researcher eventually collected 342 snapshots of 171 days within six months.

Intercoder Reliability A communication major graduate and the researcher coded 50 snapshots, about 15 percent of the overall sample, for the intercoder reliability test. According to Holsti's formula⁴⁵, the simple agreement was 80.0 percent for *Story Topic*, 95.1 percent for *Geographic Bias*, 98.2 percent for *Number of Hyperlinks*, and 86.7 percent for *Uses of Multimedia*. The researcher coded the overall 342 snapshots for the population parameter for six months of NYTimes.com.

Multiple Sampling

To compare sampling methods, the researchers divided the 342 snapshots into two individual observations—noon and evening—based on the time when the snapshot was

⁴⁵ Ole R. Holsti, *Content Analysis for the Social Sciences and Humanities* (Reading, MA: Addison Wesley, 1969).

taken. By using noon and evening observations, the researchers, in effect, manipulated two sets of data so that the sampling comparison will obtain better validity. Simple random samples of size three, four, five, and six days of snapshots were drawn from the noon observation and the evening observation of the site. Fifty samples were drawn for each sample size and each observation. Therefore, a total of four sets of 50 samples were chosen for each observation.

Sample Size Comparison

The sample means and standard deviations 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 deviations of the sample means. The Central Limits Theorem predicts that the sample means distribution is close to a normal curve and the mean of the samples means is approximately the population mean. Thus, in 95 percent of samples, the population mean should be between two standard deviations of the sample mean, and in 68 percent of samples, the population mean should fall within one standard deviation of the sample mean.⁴⁶ 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.

⁴⁶ Lacy, Riffe, Stoddard, Martin, and Chang, "Sample Size For Newspaper Content Analysis in Multi-year Studies;" Riffe, Lacy, Nagovan, and Burkum, "The Effectiveness of Sample and Stratified Random Sampling in Broadcast News Content Analysis."

Results

The Population

Table 1 shows the population means, standard deviations, and coefficient of variation for the 171-day population. During the six-month period of time, from July 6, 2005 to January 5, 2006, there were approximately 5.25 U.S. stories ($SD = 1.538$) published on the top headline portion of NYTimes.com front page every day. Meanwhile, only about 32.83 percent of the top headlines ($SD = .159$) cover the rest of the world. Among the everyday top headlines, coverage of war, defense and terrorism accounts for 12.0 percent ($SD = .100$). On average, there are 22.85 hyperlinks ($SD = 4.586$) pointing to other Web pages and files, among which 14.02 ($SD = .062$) percent are multimedia information 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 Uses of Multimedia, July 6, 2005 to January 5, 2006.

	Population Mean	Population Standard Deviation	Coefficient of Variation
Number of U.S. News	5.25	1.538	.293
Percentage of Foreign News	.3283	.15938	.485
Percentage of War News	.1200	.10026	.836
Number of Hyperlinks	22.85	4.579	.200
Percentage of Multimedia	.1402	.06179	.441

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 research usually examines coefficient of variation to

test the variability assumption and detect the impact of variability on sampling size.⁴⁷ In particular, if the coefficient of variability exceeds .5, researchers advise increasing the size of the sample.⁴⁸ The coefficient of variation for the percentage of war coverage was the highest (.836) compared to the other four variables, according to **Table 1**.

The samples

To answer the research questions and find an effective and efficient sample size, comparisons of multiple sampling predictions with the population parameter are conducted as shown in **Table 2**. According to the Central Limits Theorem, 95 percent of random sample means will be within plus or minus two standard deviations of the population mean, and 68 percent will be within plus or minus one standard deviation 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 days, the random selection of six days is shown to be the most effective and efficient sample size. Three-day sampling is apparently insufficient because using three-day sample 11 measurements fail to generate a sample mean within one standard deviation to the population mean by the chance of 68 percent, or within two standard deviations by the chance of 95 percent. Randomly selecting four days has greatly improved the sample efficacy, but out of 50 sets of samples only 92 percent in the noon observation and 94 percent in the evening observation for the *percentage of war coverage* variable are within two standard

⁴⁷ Lacy, Riffe, Stoddard, Martin, and Chang, "Sample Size for Newspaper Content Analysis in Multi-year Studies;" Riffe, Lacy, and Fico, *Analyzing Media Message*; Riffe, Lacy, Nagovan, and Burkum, "The Effectiveness of Sample and Stratified Random Sampling in Broadcast News Content Analysis."

⁴⁸ Riffe, Lacy, and Fico, *Analyzing Media Message*.

deviations; 94 percent in the noon observation and 86 percent in the evening observation for the *number of hyperlinks* variable are within two standard deviations; 94 percent in the noon observation for the *number of U.S. news*, and 92 percent in the evening observation for the *percentage of multimedia* are within two standard deviations. Meanwhile, drawing five random days seems fairly effective to estimate the six-month population except for 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 Deviations 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 Uses of Multimedia, Observations during July 6, 2005 to January 5, 2006.

Sample Size	Six		Five		Four		Three	
	1 SD	2 SD	1 SD	2 SD	1 SD	2 SD	1 SD	2 SD
Number of U.S. News								
Noon Observation	98%	100%	86%	98%	80%	94%	76%	90%
Evening Observation	98%	100%	92%	96%	76%	96%	76%	90%
Percentage of Foreign News								
Noon Observation	92%	100%	90%	96%	80%	98%	78%	92%
Evening Observation	94%	98%	88%	98%	86%	98%	76%	88%
Percentage of War News								
Noon Observation	98%	98%	80%	96%	82%	92%	74%	90%
Evening Observation	84%	98%	84%	98%	80%	94%	66%	92%
Number of Hyperlinks								
Noon Observation	88%	98%	80%	94%	72%	94%	70%	86%
Evening Observation	86%	98%	86%	96%	78%	86%	66%	94%
Percentage of Multimedia								
Noon Observation	94%	100%	84%	100%	86%	98%	86%	96%
Evening Observation	90%	98%	84%	88%	88%	92%	86%	92%

Note: The crossed percentages indicate the sampling did not meet 68% or 95% critical values.

The criteria predated by the Central Limits Theorem are not fully met until the sample size is enlarged to six. Therefore, a sample size of six days will effectively and efficiently represent the content on the updated daily news sites in a six-month period of time.

Conclusion & Discussion

Although the sample size of six days is not only effective but also efficient to predict news site content in six months, sampling methods vary depending on different research questions and designs. In the comparison of sample means and population parameters for five variables, the *percentage of foreign news* meet 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* meet the criteria in six-day and five-day samples. However, some variables like *the use of multimedia* seem to be much more "sensitive" than others in the sampling tests because among 50 sets of the five-day samples only 88 percent sample means fall within the two standard deviations 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 unaware of the "sensitiveness" or variability of the variables.

Research questions and research designs are diverse in many ways. In terms of statistical discussions, researchers might need to pay attention to testing theoretical assumptions and discovering variables' statistical features. For instance, previous sampling research suggests increasing sample size if the coefficient of variation is greater than .50. In the present study, however, more troubles are caused by variables with relatively low variability rather than those with high variability. A similar pattern was found in previous studies as well. Lacy and colleagues identified one set of samples that

failed to meet the criteria as the one with the least amount of variation in exploring seven-constructed-week sampling for multi-year studies.⁴⁹

These inconsistent findings implied 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.⁵⁰ Vogt also pointed out that the sample means distribution would be much closer to a normal curve if the sample size was 30 or more.⁵¹ With a larger sample size, the feature of a normal distribution would be more apparent. In this sampling study, the 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.

While this study investigates different simple random samples, it does not test different types of sampling methods, such as the stratified sampling that some Internet studies have used. Part of the reason is that the efficiency of simple random sample of six days has already been approved. There is no need to employ a constructed week sample to predict six months' content for the reason that one-week consists more than six days. Due to the limit of data, this study can only create a population parameter for six-month period of time. Future studies must extend the time frame to a year or multiple years,

⁴⁹ Lacy, Riffe, Stoddard, Martin, and Chang, "Sample Size for Newspaper Content Analysis in Multi-year Studies."

⁵⁰ Arthur Aron and Elaine N. Aron, *Statistics for the Behavioral and Social Sciences: A Brief Course*, 2nd ed. (Upper Saddle River, NJ: Prentice Hall) 2002.

⁵¹ W. Paul Vogt, *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*, 3rd ed., (Thousand Oaks, CA: Sage) 2005.

when a larger sample size would be required. Then a more complex sampling comparison, including the contrast of different types of sampling strategies, should 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.

Furthermore, by differentiating types of online media outlets, this study can only focus on one type of news site, which owns an independent newsgathering and editing system and is updated regularly. In addition to Web sites like NYTimes.com, further studies could replicate and advance this multiple sample size comparison on more news sites of the same type. And other types of Web sites, such as Web blogs, need more creative approach to test the sampling size effectiveness and efficiency.

Appendix

Online Sampling Coding Sheet

1. Web site: NYTimes.com
2. DATE MM/DD/2005
3. TIME: _____

4. Story topic
 - ___ 1) Politics and government acts
 - ___ 2) War and defense
 - ___ 3) Diplomacy & foreign relations
 - ___ 4) Economic activity
 - ___ 5) Agriculture
 - ___ 6) Transportation and travel
 - ___ 7) Crime
 - ___ 8) Public moral problems
 - ___ 9) Accidents and disasters
 - ___ 10) Science and invention
 - ___ 11) Public health and welfare
 - ___ 12) Education and classic arts
 - ___ 13) Popular amusements
 - ___ 14) General human interest
 - ___ 15) Others _____ (SPECIFY)

5. Geographic bias
 - ___ 1) U. S.
 - ___ 2) U.S. neighbors
 - ___ 3) Central/South America
 - ___ 4) Western Europe
 - ___ 5) Eastern Europe
 - ___ 6) Middle East and North Africa
 - ___ 7) Africa (Sub-Sahara)
 - ___ 8) South Asia
 - ___ 9) Japan
 - ___ 10) Other East Asia
 - ___ 11) Oceania
 - ___ 12) Others _____ (SPECIFY)

6. Number of links _____
7. Uses of Multimedia
 - ___ 1) Images (including slideshows)
 - ___ 2) Video clips
 - ___ 3) Audio clips
 - ___ 4) Interactive media (e.g. Flash)
 - ___ 5) Others _____ (SPECIFY)