

# **Publication Interval and Web Audience Estimation\***

by

Jongpil Hong  
Doctoral Candidate  
jphong@mail.utexas.edu  
Department of Advertising  
College of Communication  
The University of Texas at Austin  
Austin, Texas 78712-1092

Guohua Wu  
Doctoral Student  
mark.wu@mail.utexas.edu  
Department of Advertising  
College of Communication  
The University of Texas at Austin  
Austin, Texas 78712-1092

and

John D. Leckenby  
Everett D. Collier Centennial Chair in Communication  
john.leckenby@mail.utexas.edu  
Department of Advertising  
College of Communication  
The University of Texas at Austin  
Austin, Texas 78712-1092

Working Paper 1998

\*The authors wish to gratefully acknowledge the assistance of RelevantKnowledge (URL: <http://www.relevantknowledge.com>) for their invaluable support in the collection of the data which serve as the basis for this study.

# Publication Interval and Web Audience Estimation

## ABSTRACT

This paper examines the issue of publication interval of Web sites and its relationship to estimation of Web audience. Web audience data compiled by RelevantKnowledge were used to provide input to six reach/frequency estimation methods (Binomial Distribution, Beta Binomial Distribution, Conditional Beta Distribution, Sequential Aggregation Distribution, Dirichlet Multinomial Distribution and Hofmans Beta Binomial Distribution) and to provide a benchmark via tabulated schedules of twenty-five Web vehicles.

Results show that the models used in this study perform as well for RelevantKnowledge data as they did for either magazine or television data and for MediaMetrix and Recent Reading Web audience data. For the six models, all perform at differing levels of accuracy within acceptable limits of error.

The findings confirm that reach/frequency estimation methods can provide accurate estimates of the audiences of media schedules containing Web vehicles regardless of whether the Web publication interval definition is daily or weekly.

# **Publication Interval and Web Audience Estimation**

## **INTRODUCTION**

As the World Wide Web portion of the Internet has been increasingly used by advertisers as an advertising medium, a new question has evolved regarding the applicability of the standard operating approach in media planning as applied to this new medium. The major concern was if the methods developed in standard media types would work when used to estimate reach/frequency for the Internet.

In their pioneering study, Hong and Leckenby (1998) attempted to answer this question by collecting original Web viewership data using a Recent Reading Web audience survey. The Web audience data were used to provide input to six reach/frequency estimation methods to test the accuracy of traditional reach/frequency estimation. Test results showed that the models performed as well on Internet data as they did for either magazine or television data.

It was suggested from this study that this type of model testing should also be applied to data from other sources that may overcome some of the limitations inherent in the data using a survey method (recent reading method). It might be worthwhile to use user-centric data from sources such as MediaMetrix, The PC Meter Company, RelevantKnowledge and NetRatings to test models with a large sample base. This need led to another study that formally tested the performance of reach/frequency models using syndicated Web audience data of MediaMetrix, The PC meter Company (Leckenby and Hong 1998). The results of testing six reach/frequency estimation methods on a sample of 7,162 respondents showed all models except the binomial distribution model estimated reach/frequency within acceptable limits of error. The models

again proved to provide accurate estimates of the audiences of media schedules containing Web vehicles.

The current study extends the testing of reach/frequency estimation models with Web audience data provided by RelevantKnowledge. The objectives of the current study are twofold.

First, the study attempts to investigate how often Web sites update their Web contents. It would be important to know the publication interval of Web sites since the average number of people going to a site (a Web site's average audience) would be defined by a certain time interval set, for example, by an hour, a day, a week or a month. There is no currently accepted standard regarding this issue. The 99 sites reported in the RelevantKnowledge data set served as a basis of investigation for publication interval.

Secondly, this study tests the performance of reach/frequency estimation models with RelevantKnowledge data at two different Web publication intervals--weekly and daily. It would be appropriate to conduct such work as this using different definitions of site publication interval to determine the effect of this variable on estimation accuracy. Since the previous studies discussed above were solely based on weekly Web audience data, there is a need to test these models further to see if they would work about as well for daily audience data.

## **STUDY 1**

In broadcast, average audience of a vehicle might mean the average number of people viewing or listening to a program in a fifteen-minute period. In print, this might mean the average number of people who read or look into the publication during the publication interval. For the Web, there is no general consensus about how to define average issue and average audience of Web publications. Is this to be the average number of people going to a site in an

hour, a day, a week or a month? It may be that this definition should currently coincide with the purchase-of-unit interval which appears today to often be on a weekly basis rather than monthly but may become increasingly shortened as advertiser sophistication grows in this medium.

In order to obtain preliminary understanding of Web publication interval, the authors investigated the characteristics of Web sites reported in a data set obtained from RelevantKnowledge. Those sites could be classified into four major categories in general:

1. E-commerce -- sites that provide products, services or customer support online (for example, amazon.com and columbiahouse.com)
2. Content providers -- sites that desire to draw traffic by providing contents such as breaking news, computer news, entertainment news, or weather (for example, pathfinder.com and nytimes.com)
3. Internet service providers -- sites that provide all Internet-related services, such as access, measurement and Web hosting, etc. (for example, msn.com and conxion.com)
4. Search engines -- sites that become a portal with searchable content and information (for example, yahoo.com and netscape.com)

It was believed that Web sites in these categories might be updating their content as often as possible--daily or even hourly. It may be that Web users do not want to revisit sites that are slow to change their content. When the content of a Web site is not updated often enough, the rate of repeat visitors should decline. Marketers who want to use their Web sites to establish ongoing communication with their customers should watch user patterns and respond with refreshed content that is both current and relevant.

For online content providers, putting publications on the Web should significantly shorten the publication cycle. Emergence of online newspapers emphasized current news and

prepared online-only articles as incentives to checking a site several times each day. The promise of new online services is that they will break down the time delays that have been inherent in news delivery prior to the popularization of the Internet and bring news to the user's attention at the very moment that it is needed.

For search engine sites, they need to update their content and hyperlinks as often as possible since broken links are endemic to the search engine sites. Some research estimates the average Web page to change every 44 to 75 days. Twenty-eight percent change every 10 days and news Web sites change every few hours (Feldman 1997).

To get an empirical sense of the average publication interval of Web sites, the authors conducted a survey by sending email to webmasters of 99 sites reported in the RelevantKnowledge data set. An open-ended question asked in the survey was as follows: "We would like to know how often your Web site updates its content. For example, do you update your site by minute, by hour, by day, by week, by month or anything else?"

Among the 99 sites, 23 sites replied to the email survey. The result shows that publication interval varies greatly across sites (see Table 1). However, some preliminary conclusion can be drawn from the survey. Content providers and search engines tend to update their content constantly, even by minute, while Internet service providers are more likely to update their sites whenever it is needed. This is not unexpected because content sites should emphasize newness while Internet service providers have a consistent inventory of products and services and will not update unless there are new product/service releases or news releases. In general, however, it can be said that publication intervals of Web sites in these categories are more likely to be on a daily basis rather than weekly or monthly.

## STUDY 2

With data currently available from RelevantKnowledge, standard and non-standard media reach/frequency models developed for magazines, television and other media are to be tested for the Web medium. The main question here concerns whether or not these methods will perform accurately in estimating Web audience with both daily and weekly publication interval data.

First, a selection of models to be applied will briefly be described. Second, the development of data to test the models will be discussed. Results of the testing of these models are presented followed by some conclusions and limitation of the study.

### Models

Six models which have served as the basis for performance comparisons in magazine and TV, among others, will serve as the estimation methods in this study of Web reach/frequency: (1) Binomial Distribution (BIN); (2) Beta Binomial Distribution (BBD); (3) Conditional Beta Distribution (CBD); (4) Sequential Aggregation Distribution (SAD); (5) Dirichlet Multinomial Distribution (DMD); and (6) Hofmans Beta Binomial Distribution (HBBD).

These six models have been studied extensively and are selected to represent the spectrum of methods available for reach/frequency estimation (Chandon 1976; Danaher 1988a, b; Danaher 1989; Headen, Klomp maker and Rust, 1979; Hofmans 1969; Ju 1990; Kishi 1987; Leckenby and Kim 1992; Lee 1988; Rice 1985; Rust 1986; and Rust and Leone 1984, Kim 1994). In addition, many of these approaches have been available for use in proprietary formats for several years (Lancaster 1987; and Telmar 1980). These six models are described in the Appendix.

## Data Collection

Data were obtained from RelevantKnowledge, a Web audience measurement firm based in Atlanta, Georgia. The company adopts a user-centric tracking system by measuring Web usage activities among its panel members. The method RelevantKnowledge employs is basically very similar to that of A.C. Nielsen's Peoplemeter. It starts with a sampling of people from the universe and installs a device in each panel member's computer in order to record his/her activities with the medium. It projects an estimate regarding the universe based on the recorded data from their samples. This way, unlike the other methods, it can provide advertisers with both the demographic and the exposure information of a unique individual.

According to RelevantKnowledge, the definition of the universe of Web users is all people age 12 or older in the U.S. who have used the World Wide Web at least one time within the last month at home, work or college with Mac or PC with Windows 3.1, 95 and NT operating systems (RelevantKnowledge 1997, URL: <http://www.relevantknowledge.com/Products/definitions.html>).

The initial data set contained the records of 744 unique visitors who were measured on a daily basis during September 29, 1997 and October 26, 1997 (28 consecutive days). Those 744 unique visitors visited the top 99 sites, producing 13,039 records or visits. From these 13,039 records, only the records that related to the top 25 sites were extracted for analysis of the study. The initial sorting criterion that was used to extract the top 25 sites was the site reach for the measurement period above, i.e., four-week reach. Table 2 describes the summary of the extracted data and the demographics of the panel participants. The top 25 sites and their reach for the four-week period are presented in Table 3.

## Model Testing Procedure

Two sets of audience data were generated for the RelevantKnowledge data set. The first data set is based on weekly measurement periods of October 19, 1997 and October 26, 1997. There are 21 possible combinations of weekly measurement pair and only one set of weekly data is arbitrarily selected out of them. The total number of people in the measurement periods amounted to 111 (those respondents who were active in both weeks). For these data, the presumed "publication interval," therefore, is considered to be one week. This is consistent with the direction banner ads seem to be going in the industry. In their quest to delineate the "standards issue" in Web advertising, Novak and Hoffman recommended a week as the planning interval (Novak and Hoffman 1996).

The second data is based on daily measurement periods of September 29, 1997 and September 30, 1997. There are 28 possible combinations of daily measurement pair and only one set of daily data is arbitrarily selected out of them. The total number of people in these time periods was 121 (those respondents active on both days). It was presumed that publication interval here would be a day.

Based upon these data, average site audience, cumulative site audience, and between-vehicle duplication (cross-pairs) were defined on a weekly and daily basis using the definitions used by SMRB (Simmons Market Research Bureau 1989).

The Average Site Audience was calculated as:

$$\text{Average Site Audience} = (\text{Site viewers in Week 1} + \text{Site viewers in Week 2}) / 2$$

The Cumulative Site Audience was calculated as:

$$\text{Cumulative Site Audience} = (\text{Site viewers in Week 1 or Site viewers in Week 2} \\ \text{or Both Viewers})$$

The Between-Vehicle Duplication was calculated as:

$$\text{Between-Vehicle Duplication} = (\text{Site 1 and Site 2 Viewers})$$

These data were divided by 111 for the weekly data and 121 for the daily data respectively (measurement sample size) to provide estimates of the audience proportions which are input to the media reach/frequency estimation models. The above three media vehicle statistics for each of the 25 Web sites served as input data for the model estimations. Table 4 and Table 8 show the calculated average site audiences for the 25 sites as well as their cumulative audiences for weekly and daily data respectively.

### Benchmark Tabulations

For this study, 560 schedules of Web site vehicles were developed to show a reasonable test of the models' efficacy. Schedules were constructed by randomly selecting vehicles from the 25 sites available in the study. Forty schedules were developed for each of 14 schedule sets where a set consisted of a given number of vehicles from two to 15 vehicles. The vehicles each contained two insertions to be compatible with the tabulation and measurement systems over week 1-week 2 and day 1-day 2 time periods. The tabulated schedules were randomly selected from 2 to 15 vehicles in each schedule with 2 inserts each for total of 30 exposures maximum. This provided exposure distributions in range of size from 4 insertions total to 30 insertions total.

Tabulation involves, for each schedule, counting person-by-person exposure to each of the vehicles on each of the two measurement phases. This results in the "true" answer for the sample of exposure to the schedule vehicles over two occasions. This shows, in a two-vehicle

schedule, for example, the proportion of the sample exposed no times, one time, two times, three times or four times to the vehicles in the schedule.

The reach and frequency distributions for these schedules were then estimated using the six models. The performance of the models is assessed using the standard error criteria described below.

### Performance Evaluation Criteria

#### Definition of Error

In evaluating performances of different models, their accuracy depends partly on the manner in which error is defined in the study. In this study, two different error factors, error in reach estimation (AER) and error in the exposure distribution (APE), are adopted from previous studies (Kishi and Leckenby 1982; Leckenby and Kishi 1984). Danaher (1992) also used these definitions of AER and APE (he renamed AER as RER, "relative error in reach," and APE as EPOR, "error in exposure probabilities over schedule reach").

The error in the reach estimates for the test schedules was defined as the absolute value of the difference between the observed and predicted reach in terms of percentage:

#### Average percentage error in reach (AER)

$$AER = \frac{\sum(|o_i - e_i|/o_i)}{K}$$

where:

$o_i$  = observed reach of schedule  $i$   
 $e_i$  = estimated reach of schedule  $i$   
 $K$  = total number of schedules.

The error in the each exposure level is simply defined as the absolute difference between the observed and the estimate frequencies:

Average percentage error in exposure distribution (APE)

$$APE = (\sum PE_i) / K$$

where:

$$\sum PE_i = (\sum |o_{ij} - e_{ij}|) / \sum o_{ij}$$

$PE_i$  = percentage error in the schedule i

$o_{ij}$  = observed frequency at exposure level j of schedule i

$e_{ij}$  = estimated frequency at exposure level j of schedule i

$\sum o_{ij}$  = observed reach of schedule i

K = total number of schedules.

From the data, it was also observed that the between-vehicle duplications for the small audience-size sites were very small or zero. There were 130 "0" between-vehicle duplications out of 300 in weekly measurement data and 135 out of 276 in daily measurement data. This is probably due to the short measurement intervals and the large number of sites available for Web users to visit. As shown in Hong and Leckenby's (1998) study, these zero between-vehicle duplications were substituted by the estimates of duplication from available sample data of other vehicles. It was hypothesized that zero correlations between vehicles were due to measurement error; to use the models, these zero duplications needed to be estimated from available sample data of other vehicles. This was accomplished by using a concept developed originally by Agostini in 1961 called the Duplication Constant "K" and modified to act as a variable following Chandon (1976). [Note: Please see Hong and Leckenby's (1998) study for detailed discussion.]

## **RESULTS**

Tables 5, 6 and Table 9, 10 show two sample schedules consisting of three vehicles and eight vehicles for weekly and daily measurement period, respectively.

Tables 13, 14 and 15 show the statistics for between-vehicle (cross-pair) and within-vehicle (self-pair) duplication rates for the Web in this study and the magazines of the 1979 SMRB study, respectively. It should be noted that the within- and the between-vehicle duplication rates of weekly and daily data are quite similar to each other.

#### The Average Percentage Errors in Reach (AER)

The Average Percentage Errors in Reach for weekly and daily data are shown in Table 7 and Table 11 respectively. These errors tend to be small and comparable to those found in other media types in studies of the type developed here. Table 12 shows the results for magazine studies in New Zealand on AGB data, two SMRB magazine data studies, one SMRB television study, and two Web audience data studies. The best-performing model for weekly data in terms of AER was the BBD with an AER of 4.55%. This was followed by the HBBD at 4.67% and the CBD with 4.71% of AER. For daily data, the HBBD performed the best with an AER of 3.95%, followed by the CBD (4.75%) and BBD model (5.51%).

#### The Average Percentage Errors in the Exposure Distribution (APE)

The Average Percentage Errors in the Exposure Distribution for weekly and daily data are shown in Table 7 and Table 11 respectively. These errors are also quite comparable to those reported in other media types. This can be seen by examining Table 12 for other magazine and television study results on some of the same models studied here. For magazines, for example, the APE's ranged from 13 to 34 percent. For the Web schedules studied here, the range is between 21 percent and 33 percent for weekly data and 21 percent and 38 percent for weekly data (except the Binomial Distribution). Of the six models studied, the best-performing model for weekly data in terms of APE was the MSAD with an APE of 20.66%, followed by the CBD

at 24.17% and the HBBD with 30.50% of APE. For daily data, the CBD performed the best with an APE of 21.35%, followed by the MSAD (22.13%) and DMD (29.95%).

## CONCLUSION

This study examined the definition of publication interval of Web sites and its relationship with the estimation of Web audience. It tested the performance of existing reach/frequency estimation models for a sample of 560 Web media schedules ranging in size from two to fifteen vehicles and four to thirty insertions with both weekly and daily Web audience data. Results show that the models used in this study perform as well for RelevantKnowledge data as they did for either magazine or television data and for MediaMetrix and Recent Reading Web data. For the six models studied (Binomial Distribution, Beta Binomial Distribution, Morgenzstern Sequential Aggregation Distribution, Conditional Beta Distribution, Dirichlet Multinomial Distribution, and Hofmans Beta Binomial Distribution), all perform at differing levels of accuracy within acceptable limits of error.

These results suggest the traditional reach/frequency estimation methods could provide accurate estimates of the audiences of media schedules containing Web vehicles regardless of whether the Web publication interval is daily or weekly. The findings confirm those of previous studies that reach/frequency estimation methods perform as well on Internet data as they did for either magazine or television data.

One of the inherent difficulties in defining average audience in the traditional manner as in this study concerns the entrenched perceptions of the Web industry. Average audience of the variety here will always be a lower figure than would be found from site-centric measurement of

"visits" in some time period. But this issue has been faced many times before when a new methodology for audience measurement has been introduced to an existing, traditional medium.

It is now clear, however, that the reach and frequency distributions can be estimated quite accurately for schedules of Web vehicles regardless of publication interval. To mix these vehicles with other traditional media types and then estimate reach/frequency will need to await resolution by the industry of the problems of conceptual and technical natures.

Table 1

## Web Site Update Interval by Category

Site	Business Category	Update Interval
altavista.digital.com	search engine	daily
amazon.com	e-commerce	cannot discuss for security
cerf.net	ISP	as needed
conxion.com	ISP	as needed
demon.co.uk	ISP	as needed
digex.net	ISP	as needed
download.com	content provider	daily
erols.com	ISP	weekly
excite.com	search engine	every 2 weeks
infobeat.com	content provider	as needed
msn.com	ISP	every 20 minutes
netscape.com	search engine	constantly
nfl.com	content provider	as needed
nytimes.com	content provider	daily
pathfinder.com	content provider	constantly
quote.com	content provider	constantly
snap.com	search engine	daily
sportsline.com	content provider	constantly
sportszone.com	content provider	constantly
teleport.com	ISP	constantly
tiac.net	ISP	unsure
travelocity.com	content provider	several times per day

Table 2

## Summary of Data and Demographics of Panel Participants

Measurement Period	4 weeks (Sep. 29 to Oct. 26, 1997)
Number of Sites Measured	99
Total Traffic (Visits) to the Top 25 Sites	9,343
Total Number of Visitors of the Top 25 Sites	725
Male/Female Ratio	61.7/ 38.3
Average Age	Male: 36.3 Female: 34.7

Table 3

## Rankings of the Top 25 Sites and Four-Week Average Reach

Rank	Site	4-Week Average Reach %
1	yahoo.com	56.72
2	Microsoft.com	44.62
3	Netscape.com	41.53
4	aol.com	35.35
5	excite.com	33.60
6	Geocities.com	26.48
7	Infoseek.com	24.19
8	lycos.com	20.97
9	msn.com	20.43
10	Altavista.digital.com	16.40
11	Webcrawler.com	11.29
12	znet.com	10.48
13	hotmail.com	9.81
14	pathfinder.com	9.14
15	msnbc.com	8.74
16	four11.com	7.93
17	cnn.com	6.72
18	tripod.com	6.59
19	whowhere.com	6.18
20	usatoday.com	6.18
21	sportszone.com	6.05
22	download.com	6.05
23	att.net	6.05
24	hotbot.com	5.91
25	amazon.com	5.91

Table 4

Weekly Average Audience and Cumulative Audience of Web Sites  
for Week 3 and Week 4 (n=111)

25 Sites for Week 3 and Week 4	Average Audience (%)	Cumulative Audience (%)
yahoo.com	27.48	38.74
microsoft.com	24.32	31.53
netscape.com	21.62	27.93
excite.com	16.22	24.32
geocities.com	13.51	20.72
att.net	13.51	13.51
aol.com	9.91	15.32
lycos.com	9.01	15.32
hotmail.com	7.21	10.81
infoseek.com	5.41	9.91
webcrawler.com	5.41	8.11
msn.com	4.95	6.31
altavista.digital.com	4.50	7.21
znet.com	4.50	6.31
sportszone.com	4.05	7.21
cnn.com	3.60	5.41
msnbc.com	3.60	5.41
usatoday.com	3.60	4.5
hotbot.com	2.70	4.5
tripod.com	2.70	4.5
download.com	2.25	3.6
four11.com	1.80	2.7
pathfinder.com	1.80	3.6
amazon.com	.45	.9
whowhere.com	.45	.9

Table 5  
Sample Exposure Distribution for a Three-vehicle Schedule for Week 3 and Week 4  
(Two inserts each in tripod.com, aol.com, and hotmail.com)

	Observed	Binomial	BBD	CBD	MSAD	DMD	HBBD
# of exp.	%	%	%	%	%	%	%
0	71.17	66.36	70.20	71.69	71.25	72.54	71.47
1	18.92	28.17	22.02	18.19	18.97	16.71	20.17
2	9.01	4.98	6.07	8.96	8.84	9.90	6.17
3	0.90	.47	1.42	.97	.76	.30	1.72
4	.00	.03	.26	.18	.17	.54	.40
5	.00	.00	.04	.00	.00	.00	.07
6	.00	.00	.00	.00	.00	.01	.00
<b>Errors</b>							
AER	---	16.68	3.40	1.80	.28	4.75	.80
APE	---	47.62	23.73	3.57	1.87	14.74	18.73

Table 6  
Sample Exposure Distribution for an Eight-vehicle Schedule for Week 3 and Week 4  
(Two inserts each in hotbot.com, att.net, download.com, microsoft.com, whowhere.com,  
geocities.com, tripod.com, and infoseek.com)

	Observed	Binomial	BBD	CBD	MSAD	DMD	HBBD
# of exp.	%	%	%	%	%	%	%
0	35.14	25.86	34.47	33.03	36.70	25.89	37.64
1	25.23	36.50	30.08	28.50	22.72	36.47	27.83
2	25.23	24.15	18.52	22.11	24.15	24.12	16.81
3	7.21	9.94	9.60	10.81	9.51	9.94	9.19
4	4.5	2.85	4.42	4.07	5.04	2.86	4.67
5	1.8	.60	1.85	1.17	1.39	.61	2.22
6	.90	.10	.71	.26	.41	.10	.99
7	.00	.01	.25	.05	.07	.01	.41
8	.00	.00	.08	.00	.01	.00	.16
9	.00	.00	.02	.00	.00	.00	.06
10	.00	.00	.00	.00	.00	.00	.02
11	.00	.00	.00	.00	.00	.00	.00
12	.00	.00	.00	.00	.00	.00	.00
13	.00	.00	.00	.00	.00	.00	.00
14	.00	.00	.00	.00	.00	.00	.00
15	.00	.00	.00	.00	.00	.00	.00
16	.00	.00	.00	.00	.00	.00	.00
<b>Errors</b>							
AER	---	14.31	1.03	3.25	2.41	14.26	3.82
APE	---	28.89	22.56	18.12	11.43	28.86	22.11

Table 7

Summary of Average Error Calculations for Week 3 and Week 4  
Web schedules (n=560)

Model	Error Type	
	AER(%)	APE(%)
Binomial	20.54	46.78
BBD	<b>4.55</b>	32.41
CBD	4.71	24.17
MSAD	4.95	<b>20.66</b>
DMD	9.82	32.85
HBBD	4.67	30.50

Table 8

Daily Average Audience and Cumulative Audience of Web Sites  
for Day 1 and Day 2 (n=121)

24 Sites for Day 1 and Day 2*	Average Audience (%)	Cumulative Audience (%)
yahoo.com	31.40	38.84
microsoft.com	27.27	32.23
netscape.com	18.60	26.45
msn.com	12.81	15.70
excite.com	10.74	15.70
aol.com	9.50	14.88
hotmail.com	8.26	9.09
altavista.digital.com	7.44	13.22
att.net	5.79	6.61
lycos.com	5.79	9.92
usatoday.com	4.55	6.61
webcrawler.com	4.55	6.61
cnn.com	4.13	4.96
geocities.com	4.13	7.44
sportszone.com	3.31	4.13
msnbc.com	2.48	4.13
pathfinder.com	2.48	4.13
whowhere.com	2.07	4.13
amazon.com	1.65	3.31
download.com	1.24	1.65
four11.com	1.24	2.48
hotbot.com	1.24	2.48
zdnet.com	1.24	2.48
tripod.com	.41	.83

\* "infoseek.com" site is deleted because average audience was "0"

Table 9  
Sample Exposure Distribution for a Three-vehicle Schedule for Day 1 and Day2  
(Two inserts each in tripod.com, microsoft.com, and msnbc.com)

	Observed	Binomial	BBD	CBD	MSAD	DMD	HBBD
# of exp.	%	%	%	%	%	%	%
0	65.29	52.96	60.92	65.50	66.07	66.41	65.69
1	12.40	35.51	24.41	11.26	10.34	11.46	18.85
2	19.83	9.92	9.78	21.34	21.21	19.67	8.64
3	1.65	1.48	3.54	1.50	1.98	.48	4.13
4	.83	.12	1.08	.40	.40	1.91	1.84
5	.00	.00	.25	.00	.00	.00	.69
6	.00	.00	.03	.00	.00	.07	.17
<b>Errors</b>							
AER	---	35.52	12.59	.61	2.25	3.23	1.07
APE	---	97.70	70.53	9.33	12.13	9.85	63.38

Table 10  
Sample Exposure Distribution for an Eight-vehicle Schedule for Day 1 and Day2  
(Two inserts each in aol.com, whowhere.com, tripod.com, msnbc.com, pathfinder.com, att.net,  
webcrawler.com, and lybos.com)

	Observed	Binomial	BBD	CBD	MSAD	DMD	HBBD
# of exp.	%	%	%	%	%	%	%
0	61.16	50.89	61.97	62.05	63.61	59.16	61.94
1	19.83	35.11	21.81	20.31	17.30	20.99	21.84
2	13.22	11.36	9.24	11.32	11.76	16.59	9.25
3	3.31	2.29	4.03	4.07	4.84	1.39	4.03
4	2.48	.32	1.74	1.66	1.83	1.70	1.74
5	.00	.03	.73	.45	.50	2.15	.73
6	.00	.00	.30	.12	.12	.14	.30
7	.00	.00	.12	.02	.02	.00	.12
8	.00	.00	.04	.00	.00	.00	.04
9	.00	.00	.01	.00	.00	.00	.01
10	.00	.00	.00	.00	.00	.00	.00
11	.00	.00	.00	.00	.00	.00	.00
12	.00	.00	.00	.00	.00	.00	.00
13	.00	.00	.00	.00	.00	.00	.00
14	.00	.00	.00	.00	.00	.00	.00
15	.00	.00	.00	.00	.00	.00	.00
16	.00	.00	.00	.00	.00	.00	.00
<b>Errors</b>							
AER	---	26.44	2.09	2.29	6.31	5.15	1.78
APE	---	52.42	22.19	11.72	17.53	19.05	22.19

Table 11

Summary of Average Error Calculations for Day 1 and Day2  
Web schedules (n=560)

Model	Error Type	
	AER(%)	APE(%)
Binomial	20.50	50.00
BBD	5.51	37.76
CBD	4.75	<b>21.35</b>
MSAD	5.71	22.13
DMD	6.78	29.95
HBBD	<b>3.95</b>	36.04

Table 12-1

AGB New Zealand 1985 Magazine Data (n=600)

Model	Error Type	
	AER (%)	APE (%)
DMDLK	2.23	15.80
LOGLIN	1.20	12.83
CANEX	2.19	14.82
BBD	11.85	50.40

Table 12-2

SMRB US 1979 Magazine Data (n=515)

Model	Error Type	
	AER (%)	APE (%)
DMDLK	2.77	17.68
CANEX	3.08	19.91
BBD	6.46	33.23
CBD	3.12	13.76
MSAD	3.13	15.27

Table 12-3

SMRB US 1984 Magazine Data (n=508)

Model	Error Type	
	AER (%)	APE (%)
DMDLK	1.73	23.30
CANEX	3.26	25.83
BBD	4.92	33.59
CBD	3.26	17.85
MSAD	5.84	20.58

(continued)

Table 12-4

SMRB 1984 TV Data

Model	Error Type	
	AER (%)	APE (%)
ALBET	2.7	12.5
BBD-LD	3.9	5.3
DMDLK	3.5	12.6
BBD-IE	3.1	13.7

Table 12-5

Recent Reading 1998 Web Audience Data in Hong and Leckenby's Study (n=560)

Model	Error Type	
	AER(%)	APE(%)
Binomial	17.84	52.19
BBD	3.20	23.63
CBD	2.20	16.73
MSAD	3.79	26.08
DMD	3.11	38.75
HBBB	5.84	100.00

Table 12-6

MediaMetrix 1998 Web Audience Data in Leckenby and Hong's Study (n=560)

Model	Error Type	
	AER(%)	APE(%)
Binomial	21.00	41.50
BBD	2.50	8.63
CBD	2.43	9.66
MSAD	3.51	18.80
DMD	6.67	24.10
HBBB	2.23	9.78

Table 13

Average Within- and Between-Vehicle Duplication for Week 3 and Week 4

	Mean	Standard Deviation	Maximum	Minimum
Observed Within*	4.39	5.35	17.11	.0
Observed Between**	.68	1.19	9.46	.0

\* 25 Web Sites

\*\* 300 distinct vehicle pairs

Table 14

Average Within- and Between-Vehicle Duplication for Day 1 and Day2

	Mean	Standard Deviation	Maximum	Minimum
Observed Within*	4.44	6.53	23.96	.0
Observed Between**	.52	.94	7.64	.0

\* 24 Web Sites

\*\* 276 distinct vehicle pairs

Table 15

Observed and Random Within- and Between-Vehicle Duplication  
SMRB Magazines 1979 Study  
(Boyd and Leckenby 1985)

	Mean	Standard Deviation	Maximum	Minimum
Observed Within*	2.40	3.30	20.60	.20
Random Within	.40	.10	7.40	.00
Observed Between**	.30	.50	8.10	.00
Random Between	.20	.30	6.80	.00

\* n=98 magazines

\*\* n=4,753 distinct vehicle pairs

## APPENDIX

### Description of Models Utilized in This Study

#### Binomial Distribution (BIN)

Binomial distribution model constitutes the implicit landmark against which the performance of any other model should be compared since it is the simplest of all of them. The assumptions of this model are: vehicles are homogeneous; individuals are homogeneous; and vehicle exposure constitutes a set of mutually independent events such that exposure to one vehicle does not modify the probability of exposure to any other vehicle. The binomial model generally overestimates reach.

#### Beta Binomial Distribution (BBD)

This is one of the oldest known models which was developed by Richard Metheringham in the 1960's (Metheringham 1964) for use in advertising agency media planning. It is one of the simplest of the models studied and, therefore, has been frequently used in practice (Leckenby and Kim 1994). It is also the least accurate, generally speaking, of any of the known models except the binomial distribution (Leckenby and Ju 1990). But it serves as a benchmark for more complex models and may be appropriate in certain media situations. If the Internet exhibits low between-vehicle duplication, for example, then this method may be appropriate. The main problem in the BBD lies in its estimation of between-vehicle duplication (cross-pair duplication) which results in low overall estimation accuracy.

### Conditional Beta Distribution (CBD)

This method was developed by Leckenby and Kim and reported in Kim (1994) as an improvement on the traditional employment of the Beta Binomial Distribution (Metheringham 1964). In this model, it is assumed each vehicle's marginal distribution is Beta Binomial; each individual in the population is characterized by a personal probability of exposure to a given vehicle. Also, primary assumption of this model is that exposures by individuals to a given vehicle on the condition they have previously been exposed to one insertion or no insertions in this vehicle also follow a Beta Binomial. That is, the conditional distributions of exposure all follow a BBD. In addition, this model employs the Markov assumption developed by Danaher (1992) to convolute the joint exposure distribution, as estimated using the CANEX model (Danaher 1991), with each conditional distribution to form the final collapsed exposure distribution. Unlike the Danaher (1992) approach, this is a non-random convolution.

### Sequential Aggregation Distribution (SAD)

This method uses Morgenstern's reach formula to estimate reach (Chandon 1976). Then, each vehicle's marginal probability distribution is developed using the BBD separately for each vehicle. This overcomes the problem of the BBD used alone concerning between-vehicle duplication. These marginal distributions are combined sequentially to form a two-dimensional joint exposure distribution which is collapsed at each step along the main diagonals to form a marginal distribution to combine with the next vehicle's marginal distribution. It is known that different order of vehicle aggregation produces different results (Lee 1988). This method is often used in practice

(Leckenby and Kim 1994) and is very accurate (Leckenby and Ju 1990) if theoretically inelegant.

#### Dirichlet Multinomial Distribution (DMD)

The Dirichlet multinomial distribution is a member of the family of multivariate Poyla-Eggenberger distribution. It has been also called the ‘compound multinomial distribution’ (Chandon 1976) and the ‘n-dimensional basic beta distribution’ (Mauldon 1959).

Unlike the univariate distribution models, the DMD model attempts to preserve the individual vehicle homogeneity. Many of the univariate models, in contrast, average all vehicles into one univariate, composite vehicle and then treat n schedule of  $n(i)$  insertions in each of i vehicles as one vehicle with the sum of  $n(i)$  insertions.

In the DMD model, each vehicle is treated as heterogeneous and the probability of exposure to each vehicle is incorporated to obtain the exposure distribution parameters. The Dirichlet distribution provides the distribution of probabilities of individuals to be exposed to none, any one, any two, or up to all vehicles assuming one insertion in each vehicle.

#### Hofmans Beta Binomial Distribution (HBBD)

Hofmans (1969) developed a method to calculate reach for any combination of media based on Agostini’s estimation method. However, the Hofmans reach estimation is slightly different from Agostini’s in that the ‘K’ variable in his formula is used for each duplication pair instead of a constant ‘K’ as in Agostini’s formula.

Despite the superiority in theory, Chandon (1976) found that the Hofmans accumulation model did not significantly improve on Agostini's model. However, the Hofmans model for cumulative net coverage has been found to produce reach estimates better than beta-binomial distribution model (Leckenby and Boyd 1984, Kishi 1983).

In this application, after the estimation of reach is completed using Hofmans' model, a beta binomial distribution is fit to this reach post hoc using the method of means and zeros (Anscombe 1950).

## REFERENCES

- Anscombe, F. J. (1950), "Sampling Theory of the Negative Binomial and Logarithmic Distributions," *Biometrika*, 37.
- Chandon, Jean-Louis (1976), *A Comparative Study of Media Exposure Models*. Unpublished doctoral dissertation, Northwestern University.
- Danaher, Peter J. (1988a), "Parameter Estimation for the Dirichlet-Multinomial Distribution Using Supplementary Beta-Binomial Data," *Communications in Statistics*, A17, 6 (June), 777-88.
- \_\_\_\_\_ (1988b), "A Loglinear Model for Predicting Magazine Audiences," *Journal of Marketing Research*, 25 (November), 356-62.
- \_\_\_\_\_ (1989), "An Approximate Loglinear Model for Predicting Magazine Audiences," *Journal of Marketing Research*, 26 (November), 473-9.
- \_\_\_\_\_ (1991), "A Canonical Expansion Model for Multivariate Media Exposure Distributions: A Generalization of the 'Duplication of Viewing Law'," *Journal of Marketing Research*, 28 (October), 361-7.
- \_\_\_\_\_ (1992), "Some Statistical Modeling Problems in the Advertising Industry: A Look at Media Exposure Distributions," *The American Statistician*, 46 (November), 254-60.
- Feldman, Susan E. (1997), "It was here a minute ago!: archiving on the Net," *Search*, October, 9(5), 52-62.
- Headen, Robert S., Jay E. Klompmaker, and Roland Rust (1979), "The Duplication of Viewing Law and Television Media Schedule Evaluation," *Journal of Marketing Research*, 16 (August), 333-40.
- Hofmans, Pierre (1969), "Measuring the Cumulative Net Coverage of Any Combination of Media," *Journal of Marketing Research*, 3 (August), 269-78.
- Hong, Jongpil, & John D. Leckenby (1998), "Audience Duplication Issues in WWW Media Planning," In *Proceedings of the Conference of the American Academy of Advertising* (pp. 37-45). American Academy of Advertising.
- Ju, Kuen-Hee (1990), *Simple Approaches to Modeling Advertising Media Exposures*. Unpublished doctoral dissertation, The University of Texas at Austin, Austin, Texas.

- Kim, Heejin (1994), *A conditional Beta Distribution Model for Advertising Media Reach/Frequency Estimation*, The University of Texas at Austin, Austin, Texas.
- Kishi, Shizue (1987), "Exposure Distribution Models in Print, Spot-TV, and Mixed-Media Schedules: Empirical Test on Japanese Data," The Annual Conference of the European Marketing Academy, Ontario, Canada.
- \_\_\_\_\_ (1983), *Exposure Distribution Models in Advertising Media*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, Illinois.
- \_\_\_\_\_ and John D. Leckenby (1982), "A Test of the Direct/Indirect BBD and Other Exposure Distribution Models," in *Proceedings of the American Academy of Advertising*, 46-52.
- Lancaster, Kent M. (1987), "Optimizing Advertising Media Plans Using ADOPT on the Microcomputer," Paper presented at the Fourth AMA Microcomputers in Marketing Workshop, University of Hawaii at Manoa Honolulu.
- Leckenby, John D. and Heejin Kim (1994), "How Media Directors View Reach/Frequency Estimation: Now and a Decade Ago," *Journal of Advertising Research*, 34 (September/October), 9-21.
- \_\_\_\_\_ and Heejin Kim (1992), "Unresolved Issues in Media Reach/Frequency Models," in *Proceedings of the American Academy of Advertising*, 100-106.
- \_\_\_\_\_ and Jongpil Hong (1998), "Using Reach/Frequency for Web Media Planning," *Journal of Advertising Research*, January/February, 38, 1, 7-20.
- \_\_\_\_\_ and Kuen-Hee Ju (1990), "Advances in Media Decision Models," *Current Issues and Research in Advertising*, 311-357.
- \_\_\_\_\_ and Marsha M. Boyd (1984), "An Improved Beta Binomial Reach/Frequency Model for Magazines," *Current Issues and Research in Advertising*, 1-24.
- \_\_\_\_\_ and Shizue Kishi (1984), "The Dirichlet Multinomial Distribution as a Magazine Exposure Model," *Journal of Marketing Research*, 21 (February), 100-106.
- Lee, Hae Kap (1988), *Sequential Aggregation Advertising Media Models*. Unpublished doctoral dissertation, University of Texas at Austin, Austin Texas.
- Mauldon, J. M. (1959), "A Generalization of the Beta-distribution," *The Annals of Mathematical Statistics*, 30 (June), 509-520.
- Metheringham, Richard A. (1964), "Measuring the Net Cumulative Coverage of a Print Campaign," *Journal of Advertising Research*, 4 (December), 23-28.

Novak, Thomas P. And Donna L. Hoffman (1996), "New Metrics for New Media: Toward the Development of Web Measurement Standards," [URL: <http://www2000.ogsm.vanderbilt.edu/novak/Web.standards/Webstand.html>].

RelevantKnowledge (1997), [URL: <http://www.relevantknowledge.com/Products/definitions.html>]

Rice, Marshall D. (1985), *Television Exposure Distribution Models in Advertising Media*, Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

Rust, Roland T. (1986), *Advertising Media Models: A Practical Guide*. Lexington, MA: Lexington Books.

\_\_\_\_\_ and Robert P. Leone (1984), "The Mixed Media Dirichlet-Multinomial Distribution: A Model for Evaluating Television-Magazine Advertising Schedules," *Journal of Marketing Research*, 21 (February), 89-99.

Simmons Market Research Bureau (1989). *Simmons Technical Guide*.

Telmar Media Systems, Inc. (1980), "Selected Systems for Media Planning and Buying," New York: Telmar Media System Inc.