The Application Of Neyman-Pearson Methodology To The Estimation Of Web Advertising Viewers

Jess S. Boronico (E-mail: boronicoj@wpunj.edu), William Paterson University
Edward W. Christensen (e-mail: echriste@monmouth.edu), Monmouth University

Abstract

As electronic commerce continues to expand as a medium of choice in transaction markets, issues of advertising effectiveness and exposure have become increasingly salient. While the literature is replete concerning levels of exposure for more traditional forms of advertising and commerce, the extension of much of this work to the electronic commerce market is limited, despite the wealth of online tracking information that does not exist in more traditional markets. This manuscript addresses this limitation by considering the important issue of estimating advertisement exposure for "banner" or "target" advertisements within electronic commerce through the utilization of standard Neyman-Pearson statistical methodology.

1. Introduction And Literature Review

*D*espite recent market downturns in many technological sectors, the area of electronic commerce continues to witness dramatic growth. This growth is simultaneously fueled by both an increasing base of Internet users, and an accompanying increase in consumer confidence as it concerns electronic transactions. An important issue that bears exploration during this growth involves securing appropriate market share, utilizing effective advertising media, and subsequently implementing productive advertising campaigns in an environment featuring few significant consumer switching costs in addition to numerous alternative retailers.

The importance of web advertising has been well documented, and is evidenced by recent annual advertising expenditures well exceeding eight billion dollars, a significant increase from historical expenditures of $4.5 and $1.9 billion in 1999 and 1998, respectively (Saunders, 2000). Simultaneously, the cost of advertising per view/impression has decreased (Parker, 2000). These trends jointly suggest the need to evaluate the efficacy of web advertising, including the overall exposure for various forms of web marketing, embodied within one of the major challenges that electronic commerce retailers must contend with: customer acquisition (Hoffman and Novak, 2000).

Numerous artifacts exist concerning the optimum allocation of resources and the resulting effectiveness of advertising medium for traditional commerce retailers. For example, linear programming has been widely utilized, although more contemporary perspectives include the implementation of behavioral components, including prospect theory (Berger and Smith, 1997) and the use of optimal control theory (Narasimhan and Ghosh, 1994). Despite these efforts, measuring advertising effectiveness remains an inexact science for traditional campaigns and is typically retrospective with little information concerning the actual number of individuals who have viewed an ad, acted upon that ad, and made a purchase based upon that advertisement. However, current technological media now allow for a more careful inspection and monitoring of advertising campaigns. For example, common measurement for technologically based alternatives include hit counts, impressions, click-throughs, and online purchases, among others. Of considerable importance, and the focus of this manuscript, is the number of impressions (e.g. views of the advertisement) that are obtained from any advertising outlet/sponsor: Costs for advertising medium are often stated in terms of "cost per thousand impressions" (CPM), and typically vary between thirty and sixty dollars per CPM (Ad Resource, 2001).

Impressions are gathered for advertising media whose advertisements are ordinarily classified as being either "banner" or "target," in style (Novak and Hoffman, 1997). The former are small rectangular graphical images, while the latter are full page advertisements, or "web-ads" (Raman and Leckenby, 1998). Each visit to a page bearing these ads is considered an "impression." Since pricing is often based on the number of impressions, the estimation of the number of impressions for a banner or target advertisement is of considerable interest for any company investing in some form of web advertising. While sponsors ordinarily provide an indication of the expected number of impressions an advertisement will make, an independent evaluation, or estimation, might assist web-based retailers in determining whether they are being over or under charged. This manuscript addressees this estimation. More specifically, it is shown how standard Neyman-Pearson statistical methodology may be implemented to assist web-based retailers in estimating the number of impressions made by either a banner or target advertisement. The implementation of this technique relies upon data that can be gathered using technological means without the risk of infringement or the utilization of proprietary data.

The manuscript proceeds as follows: section two describes the statistical technique and indicates how the technique may be implemented within a web-based environment. Section three presents a few limitations, with section four providing an illustrative numerical example. Section five is by way of final comments.

2. Impression Estimation: A Statistical Approach

As noted earlier, a fundamental concern of the effectiveness of a web-based advertising campaign involves an accurate estimation of the number of viewers, or impressions, generated by the advertisement, by visiting a sponsor's web page. This number of "impressions" may be estimated through the procedure discussed below, and involves the collection of two equal sized samples, drawn on subsequent days, and the application of statistical methodology.

The particular technique suggested here for estimating this number of impressions for a web-based advertisement is based on Neyman-Pearson statistical methodology and follows a format that is often used environmentally in estimating an endangered specie's population size. For example, the Ecological Society of America (ESA) utilizes the methodology presented here for estimating the risk of extinction (Commission on Life Sciences, 1995), with a similar approach suggested by Taylor and Gerrodette (1993). In general, the size of a population is estimated through multiple sampling, where elements of the first sample are marked. The subsequent second sample, which is equal in size to the first marked sample, utilizes the percentage of elements marked from the first sample and captured within the second sample, in order to provide the resulting estimate for the total population size. This population estimate would then be used as an approximation to the number of viewers/impressions generated from a sponsor's web page, and have exposure to an advertiser's banner or target advertisement.

More specifically, as it applies to impressions for a web-based advertisement, an advertiser would be interested in measuring the total number of viewers for a potential advertisement (N) placed on the sponsor's web page. To estimate this parameter the advertiser would first measure the number of users who viewed the advertisement on the sponsor's web page and subsequently visited their own web page (i.e. the advertiser's page) by "clicking through" from the advertisement placed on the sponsor's web page during time/sampling period 1 (n). To ascertain this value of n, the advertiser's web server log may be searched, over sampling period 1, for those individuals that visited the advertiser's web page from the referring URL representing the sponsor's web page. The simultaneous unique identification of the potential customer's IP address and the machine name for the user may then used to filter out these unique users and consequently "tag" these users as having visited the advertiser's web page from the advertisement placed on the sponsor's web page during sampling period 1.

The advertiser then must proceed to collect data over a second, non-overlapping sampling period, designated simply as sampling period 2. During this second sampling period, the advertiser must determine the number of click-throughs from the sponsor's page to the advertiser's web page who have also been tagged as a previous visitor from the sponsor's web page during the previous sampling period. In other words, the advertiser must identify the number of "repeat" click-throughs from the sponsor's page to their page, or those visitors that clicked through during both time periods (k).

To determine *k*, the web server log is again searched for those individuals visiting the advertiser's web page, during the second sampling period, from the sponsor's referring URL. For these visitors whose referring URL during time period 2 is the sponsor's URL, the visitor's IP address is again noted, and if this visitor has also accessed the advertiser's web page from the same URL during sampling period 1 (and hence was tagged during period 1), then a "match" is identified. The total number of matches found during sampling period 2 represents *k*. The number of visitors sampled during the second sampling period, from which *k* is determined, is based only from those visitors whose referring URL during the second time period is the sponsor's URL. That is, sampling during period 2 continues until *n* (the number of tagged click-throughs gathered during time period 1) visitors have been found with the sponsor's referring URL within the second sampling period. It should be noted that the unique identification of users can also be accomplished, in addition to utilizing a web server log, via data captured by a third party (click-through tracking), the use of a cookie, or user login tracking. Particulars concerning the efficiency embodied by each method is left as an implication for future research, although we briefly comment, below, on these alternatives:

Web server logs: Communication between a web browser and a web server results in an entry in the web server's log recording a "transaction." The data that is actually maintained in this log file varies according to the type of server being used and the log file format supported. Two commonly employed formats are the "common log file" and "extended log file" formats, both of which contain the IP address of the requesting computer/visitor, the date and time of the request, the originating or referring URL, the protocol used for the request, and the browser and operating system used by the requesting computer.

Tracking users via the web server log is limited to the extent that each IP address does not necessarily identify a unique user. For example, when employing "dynamic" IP addresses, a particular machine or user is assigned a different IP address each time they log in, thereby eliminating the unique one-to-one correspondence between user and IP address. Consequently, multiple users may have the same IP address over time and can not be distinguished as unique users by the web server log. This limitation may be somewhat moderated, in that domain names for particular IP addresses can be resolved through a reverse DNS lookup, thus making it possible to identify the general class of users that might incur this "multiple IP address" limitation. For example, it is possible to identify AOL, government, or nation specific users.

Click-through tracking: Users, as they click on an advertisement, have information transferred to a click-through database which logs the request for the advertiser's page and then returns a redirect command to the client thus directing the user to the desired advertiser's page. This technique captures much of the same information as described above, however, it allows for the same advertisement to appear in multiple locations and is managed by a third party.

Cookie tracking: When a user accesses a web page the server places a small data file (cookie) to the user's hard disk for tracking purposes (Peters and Sikorski, 1997). A variety of information, similar to that discussed earlier, is stored in the cookie. The use of cookies for the purpose presented here has limitations as it associates a user with a specific computer. Consequently, a user visiting a site from multiple locations may be omitted since the information stored in the two cookies will not match.

User login tracking: Perhaps the most robust method of tracking involves acquiring information from the user through a unique login/password allocation system that is used to access the site from any location. One potential limitation, however, involves a user sharing a password/login with another user thus compromising this unique identity. The further potential associated with the inconvenience of logging on to the web page may also compromise potential business. However, if a web retailer chooses to adopt the statistical technique presented here in order to measure exposure to a more traditional form of advertising, such as a newspaper ad, where there is no referring URL or user IP address, then this form of tracking may form the only viable alternative.

Once both *n* and *k* have been established, the following equation may be utilized for estimating the number of advertising views/impressions (*N*) , at an ex-ante declared level of statistical confidence (α):

$$\left| \frac{\dfrac{k}{n} - \dfrac{n}{N}}{\sqrt{\dfrac{\dfrac{n}{N}\left(1 - \dfrac{n}{N}\right)}{n}}} \right| \le z_{\alpha/2} \tag{1}[1]$$

where $z_{\alpha/2}$ represents the standard normal variate. The resulting range for *N* may then be used as an estimate for the number of views/impressions generated by the sponsor's web page.

3. Modeling Limitations And Extensions

In addition to the tracking implications discussed earlier, the preceding technique is subject to additional statistically oriented limitations within the context described here. For example, it is assumed here that (a) the population of visitors to a sponsor's web page is relatively stable over the time that the samples are taken. In other words, the number of visitors to the sponsor's web page during any given twenty four hour window remains stable over time. While ordinarily this might be true, there does exist the possibility that the number of visitors to a sponsor's web page may show significant variability from day to day. Second, and perhaps more critical, is (b) the requisite independence assumption that any user's propensity, or likelihood, to click through to the advertiser's web page from the sponsor's page remains constant. This assumption may not be robust as it concerns web advertisements, since a web user's experience during the first access to a web page may impact on the probability that the page is accessed when visiting the sponsor's page during the second day. We should note however, that users who decide to visit the advertiser's page directly during the second day, or access the advertiser's web page from an alternative URL during the second day, rather than linking directly from the sponsor's page, do not impact on the precision of the methodology presented here. Also, to moderate the potentially debilitating effect of violating this independence assumption, the web advertiser may place an alternative advertisement during the second sampling period: the two advertisements might address the same thematic elements but be differentiated to the extent that the identification of the specific advertiser is not transparent. This would increase the likelihood that a visitor at the sponsor's page might click thorough to the advertiser's page on both days in an independent manner. The specific impact of the violation of either of the assumptions discussed here is not pursued, but is left as a statistical implication for future research.

Related to the limitations noted above, it is also important to note that the model may be best adapted to sponsor sites that draw a heavy volume of repeat users. This is because if an impression is made on day 1 and this visitor clicks-through to the advertiser's site, we wish for the population of users on day 2 to contain many of the same users from day 1. An extreme case will demonstrate this importance: If a sponsor's site draws customers who are not inclined to visit again immediately, then any click-through to the advertiser's site may simply be bookmarked by the user for future reference, and the advertiser will not have the opportunity to capture this user again on day 2 since the user does not intend to return to the sponsor's site so quickly. Since the method relies on capturing two samples from the same population, this extreme case implies that there may be no users who are captured on both days. The resulting population estimate, in this case, will turn out unrealistically large, where, in fact, the number of impressions is actually much smaller than this large estimate: this discrepancy would emerge from the fact that none of the population of users return to the sponsor's site on the second day.

Lastly, it should be noted that the methodology presented here can also be applied to more traditional media, such as newspapers. Once again, visitors who access the advertiser's site would need to be tagged, but since they are not click-throughs (having gone to the advertiser's site by virtue of the written print ad), they would need to be tagged by using an alternative mechanism. For example, visitors during sampling period 1 could be queried for

---

[1]   William L. Quirin (1978), *Probability and Statistics*, pp. 321-322.

information including where they were exposed to the advertiser's web page. If they indicate that they were directed to the advertiser's web page from the newspaper advertisement then they could be provided with a unique password (thus being tagged). On the second day an alternative advertisement could be placed in the same location, in the same newspaper. Visitors during the second sampling period could again be queried, and asked to provide their password from the first day. Only those visitors who have previously been tagged (by providing the password received during sampling period 1) would be considered as part of the viable sample for period 2, and from these, the group that indicated that they have also seen the advertisement on day 2 (perhaps they could be asked to enter a password shown in the day 2 advertisement) are counted in group k. More precise sampling methods are left as an implication for future research.

4. Illustrative Example

The AD corporation has placed an advertisement on Company SP's web page. The marketing and sales departments of AD have agreed that this advertising media is worth utilizing if the CPM (cost per thousand daily exposures) is approximately $25. A cost higher than this implies that the media alternative is too expensive for AD. Since the cost of running a banner ad on SP's page is fixed at $100 per day, AD would like to know how many exposures, or impressions, are achieved from this banner ad. Hence AD would like to estimate the number of viewer's to SP's web page during the day (N).

The information systems unit at AD is informed that they are to check the web server log between noon on May 1 and noon on May 2 (sampling period 1) in order to ascertain how many visitors to AD's web page have a referring URL that matches that of SP's web page (i.e. click-throughs). They are also to log the IP address for all these users. At the end of the day they are to report the number of users who have visited AD's web page from the referring page (SP's). This value represents n. Following the completion of this sampling period, the information systems department reports that there are $n = 750$ such click-throughs.

The information systems department is now informed that they are to follow a similar procedure the following day. That is, they are to again check the web server log, beginning at noon on May 3, for those visitors that have a referring URL matching the sponsor's web page (or, for those users who do not have the sponsor's referring URL, there is a matching IP address to one of the tagged users identified in sampling period 1). For each of these click-throughs from the sponsor's URL they are to again check the IP address for this user. If the IP address matches one of the IP addresses that has been previously logged from sampling period 1, this is considered a "match," representing a repeat user. The total number of such matches is to be reported at the end of this second time/sampling period. The second sampling period does not necessarily end at noon of the following day, but finishes when there are $n = 750$ users identified at AD's web page with the referring URL matching SP's web page (or for those users not coming from the referred URL having a repeat IP address). The information systems department reports that 700 such click-throughs have been identified at 1:35 PM on May 4, hence sampling ends at that time. The information systems department also reports that during this second sampling period (noon May 3 – 1:35 May 4), there were $k = 133$ IP addresses that were matches during the first and second sampling periods.

The marketing department now employs equation (1), and substitutes the relevant values at an $\alpha = .90$ level of significance:

$$\left| \frac{\frac{133}{700} - \frac{700}{N}}{\sqrt{\frac{\frac{700}{N}\left(1 - \frac{700}{N}\right)}{700}}} \right| \leq z_{\alpha/2} = 1.65$$

Solving for *N* results in the following: $N \in [3246, 4197]$. Hence the conclusion is that the number of viewers, or exposures, to SP's web page during the day is between 3246 and 4197. Since the cost to run the banner ad, per day, is $100, the CPM is quickly calculated over this range of values for *N*. Cost per exposure $\in (\$308, \$238)$. Since the target CPM, $25, is within this interval, AP is comfortable with continuing to use SP's web page as a media alternative and placement of a banner advertisement.

5. Concluding Comments

As the purging and clearance of inefficiencies in the "dot-com" market continues to prevail, advertisement value and the effectiveness of technological marketing tools must remain under close scrutiny. Given the proprietary nature of much of the data found in the technology market, it is important that potential retailers consider numerous techniques designed to measure the efficacy and cost-benefit ratio for these marketing alternatives utilizing data that is available and at their disposal. The suggested methodology presented here offers one such alternative, where the data required often resides on the company's server.

The methodology presented here, while practical, does offer implications for future research, including the refinement of data collection techniques and methodology to address the limitations cited earlier, and the extension to more sophisticated forms of statistical sampling and data collection. However, given the increasingly diverse and global nature of markets and the growing reliance on e-commerce and increasing popularity of web-based retailing, the continued exploration of statistical techniques to evaluate the relative advantages and disadvantages of alternative marketing alternatives, not only comparing contemporary technological alternatives to more established alternatives (radio, television, newspaper), but contemporary technological alternatives to themselves (i.e. a comparison of alternative sponsor sites) will help retailers address the challenges brought about through an increasingly competitive marketplace requiring effective cost maintenance and dynamic marketing strategy. 🔖

References

1.   Ad Resource(2001) "Sample ad rate guide", *http://adres.internt.com/adrates/print/0,,9251_198601,,0.html*.
2.   Berger, P. D. and G. E. Smith "The impact of prospect theory based framing tactics on advertising effectiveness", *Omega*, 26 (5), 593-609
3.   Hoffman, D. L. and T. P. Novak (2000) "How to acquire customers on the web", *Harvard Business Review*, 78 (3), 179-88.
4.   Narasimhan, R. and S. Ghosh (1997) "A dynamic model of manufacturing quality's effect on optimal advertising and pricing policies", *European Journal of Operational Research*, 72 (3), 485-502.
5.   Novak, T. P. and D. L. Hoffman, D. L. (1997) "New metrics for new media: Toward the development of web measurement standards", *World Wide Web Journal*, 2(###), 213-46.
6.   Parker, P. (2000) "Report: CPMs soften, e-commerce fuels online ad growth", *Internet News Advertising Report*, *http://www.internetnews.com/IAR/print/0,,12_29774,00.html*.
7.   Peters, R. and R. Sikorski (1997) "Cookie Monster?", *Science*, 278(###), 1486-87.
8.   Raman, N. V. and J. D. Leckenby (1998) "Factors affecting consumers' Web-ad visits", *European Journal of Marketing*, 32 (7/8), 737-48.
9.   Saunders, C. (2000) "Numbers show signs of online ad slowdown", *http://cyberatlas.internet.com/markets/advertising/print/0,,5941_542091,00.html*.