Correcting Bias Caused by Prototype Use Response Heaping

Abstract

Imprecision in recall resulting in response heaping occurs, often by supplying a value central to a range of values (i.e., by prototype use). For natural resources, heaping can be for days spent (e.g., hunting), animals/fish harvested/seen or amount spent. Frequency distributions having heaps can cause significant bias in such measures as means or totals (e.g., of days of hunting or amount harvested). Since bias can cause poor decisions, determining if bias is large enough to matter is important. The paper provides the logic and flow of a deheaping procedure allowing determination of bias in means and totals for prototype heaping. The SAS© program developed and available on the web is for heaping at multiples of 5 and 7. Deheaping and its consequences are illustrated. The article discusses variables being appropriate to be deheaped by the program, why only multiples of 5 and 7 are covered, bias other than from prototype use and reasons for using the program rather than alternatives.

Introduction

A number of papers in the human dimensions, wildlife and tourism literature involve analysis of variables that have excessive frequencies occurring for some numbers such as on numbers ending in zero or five. This paper is about deheaping when heaping results from using a value as a proxy for a range of values when there is uncertainty about an exact response to use. Huntenlocher, Hedges & Bradburn (1990) refers such responding as using prototypes. A respondent who is not confident, for example that 23 is the number of days hunting moose in a year may give the prototype response of 25. In this context a prototype used by many respondents could be "if I'm not confident about a response but feel my response should be between 23 and 27, I'll respond 25." A less confident respondent might "round" 23 to 20 and 26 to 30. The first respondent can be referred to as a 5-heaper and the second as a 10-heaper. Deheaping refers to starting with frequency data such as f (x) in Figure 1, and arriving at a curve such as d(x) in which responses in heaps have been allocated back to give the distribution heaped responses likely came from.

A manager might want to use data on total person-days of hunting or amount of harvest in seting season length and / or allowable harvest. If the manager or researcher see heaps at responses such as 5, 10, 15, 20, etc. she/he may feel confident that the heaps are occurring because of using a "rounded" response when uncertain about a value to give. Use of "rounding" could seem obvious because no licenses or hunting limits would result in excesses of responses ending in zero or five. Chance would not give preference to any particular number so peaks, other than from random variation are not expected. Unfortunately, heaping can cause bias in estimates that greatly exceeds estimated standard deviations in estimates such as means or totals (e.g., mean days of hunting a species such as moose or in the mean or total harvest of ducks or fish).

Literature

Consideration of the implications of heaping in natural resources data has a history. Beaman, Vaske and Grenier (1998) presents a deheaping methodology. Vaske and Beaman (2006) provide a review of findings on response heaping by 2006. Here, only literature directly relevant to deheaping is cited.

Huttenlocher, Hedges and Bradburn (1990) deal with respondents giving imprecise responses using prototypes. One can think of a prototype as a triple, [L,V,U]. In the triple, here enclosed in square brackets, L is the lower limit of a range, U is the upper bound of a range and V is the prototype value given when a person feels uncertain about x with $L \leq x \leq U$. Because of uncertainty about a value, the person only feels comfortable giving V as a response. In other words, using a prototype refers to a person using a particular response, say 20, when the person feels some value, say between 18 and 22, is most likely the right answer to give but does not feel comfortable giving a particular value. In their thinking V is not a precise value but a good approximation. In this context, some respondents give precise responses and others give imprecise, prototype, responses (e.g., respond using responses that are multiples of 5 or 10).

Beaman, et al. (1998) use assumptions in justifying a procedure for deheaping when heaping results from the use of prototypes. General ideas behind their deheaping can be understood by looking at Figure 1. The frequency function, f(x), has heaps. Below the heaps one can discern a relatively smooth trend function, t (x), that traces the trend in frequency of non-heap responses. Since the trend establishes where heaps start, an estimate of the number of responses in a heap at x is h(x)=f(x)-t(x).

Deheaping involves arriving at a function, δ(x), showing the way that a frequency function would look if all respondents gave precise responses; that is did not use prototypes or other response strategies such as multiples that result in heaps (Vaske, Beaman and Huan article re multiples- reference not handy). Therefore knowing the size of heaps, one just needs to prorate the number of responses in heaps appropriately to x-values they came from. The Beaman, Vaske and Grenier (1998) assumption that prorating back should be roughly in proportion to the t(x) values in the range of a prototype, allows prorating back. The idea is that for the prototype [L,V,U], Equation 1 defines the proportion that goes to x. In words, the part of $H_{[L,V,U]}$ that goes to x is in proportion to t(x) as a part of Σt(j) over the range of the prototype. This approach is consistent with the mathematical statistics concept used in addressing the consequence of grouping, you could say heaping, responses that have a given distribution (e.g., see Yates 1934).

Equation 1: $P_t(x, [L,V,U]) = t(x)/(\sum_{j=L}^{U} t(j))$

where t refers to the estimated trend below the heaps corrected for certain responses

The rationale behind the proration is easy to understand. The idea is that in the range of the prototype [L,V,U], uncertain responders have a probability Π([L,V,U]) of responding by V. Beaman, Vaske and Grenier (1998) essentially describe having such a probability as making an unbiased response. The response is unbiased in that the value of the precise responses that a person would give does not influence using the prototype. In other words, the assumption is that a person whose precise response should be L, U or any value between has the same probability of “heaping” using the prototype. Accepting this assumption is a topic in the Discussion.

Respondents using prototypes and responding in an unbiased way does not mean that estimates made using their responses are unbiased. Typical estimates made using data on days of hunting or fishing or on harvest in these days are total person-days of activity (e.g., hunting moose), total harvest (e.g., ducks harvested in a state) or means such as mean days of hunting or mean number of ducks harvested per hunter. Using [L,V,U] in a range where frequency is decreasing results in more low responders (toward L) heaping than high responders (toward U) being in a heap so treating total days or harvest for N respondents as NV is an overestimate. Respondents in being unbiased create a bias that analysis must remove based on the aggregate pattern in f(x).

Bias resulting from using prototypes is a separate matter from bias resulting from some other recall matters. Recalling the value of a variable can be subject to biases such as telescoping, bounding or forgetting, or other factors. For some respondents, the cognitive process of answering questions about their behavior involves the simple retrieval of information from memory, for example using episode enumeration (Sudman, Bradburn, & Schwarz, 1996; Nadeau & Niemi, 1995). Response errors occur from episode omission and/or misplacing episodes in time (telescoping). Episodic enumeration can give way to, for example, respondents using estimation heuristics (Tversky & Kahneman, 1974). Automatic estimation (Hasher & Zacks, 1984; Nadeau & Niemi, 1995) involves a sense of relative or absolute frequency (e.g., I go hunting once most weekends in the season). This is similar to formula based estimation such as using a multiplier (Vaske, Huan and Beaman). Responses are not necessarily correct. However, what matters for use of a prototype is whether a person is prepared to give a precise response or, because of uncertainty, gives an imprecise/approximate answer based on a prototype.

Deheaping Logic

Consider that values of a variable x are collected for a sample. For each prototype used, consider that a probability Π([L,V,U]) exists giving the proportion of respondents at any x for $L \le x \le U$ expected to heap using [L,V,U]. Given that δ(x) is the function that would occur if all respondents replied without using prototypes, the expected number in a heap from using the prototype is the number to be returned. Equation 2 shows the number in the heap. Because the total number is proportional to δ(x) at each x, the proportion to return to x is defined by Equation 3. Numbers of responses to return from the heaps to where they came from is therefore expressed by Equations 4.

Equation 2: $N([F,V,U]) = \Pi([L,V,U]) \left(\sum_{j=L}^{U} \delta(j)\right)$

Equation 3: $P_{\delta}(x, [F,V,U]) = \frac{\delta(x)}{\sum_{j=L}^{U} \delta(j)}$

Equation 4: $R(x, [L,V,U]) = P_{\delta}(x, [F,V,U])\, N([F,V,U])$

A challenge is returning numbers of respondents in heaps to x values without knowing δ(x). As long as there is a trend observable below the heaps, say t(x), the trend can be estimated by regression. Given that one does not know the functional form of t(x), assuming a form presents the risk of poor fit. Because the goal is not determining a functional form, an alternative to assuming a functional form is to use multiple overlapping regressions across the range of x. Rather than make a linear approximation over a range where there is curvature, approximating the curve in a limited range by a quadratic is an option. Using a quadratic rather than going to a cubic is reasonable given that in the range covered first and second derivatives of the function reflect most of the change taking place.

Multiple overlapping regressions can be run, for example, using SAS©``s PROC REG (SAS\STAT 1989). With f(x) set to missing for x-values that are heaps PROC REG will return estimated values for the trend at heaps as well as estimated y-values for other x-values. As indicated in Equation 5, the estimated trend, ŧ(x), can be defined as the weighted average of estimates that occur for each x value. Given, for example, that every regression is run with at least six non missing f(x) defining the trend (more than 6 only to avoid leaving out f(x) values near the maximum of x), it is reasonable to have weighting for estimates that are lower for first and last points in the range of the regression so big swings that can occur at the ends of ranges would not unduly influence trend values. An alternative that may be pursued in modifications to the program is to use estimated variance in y-values produced by the regression program. That was not pursued since these values can be highly variable with three parameters being estimated based on 6 observations (i.e., these variance estimates can be very inaccurate).

Equation 5: ŧ(x)=the weighted average of regression estimates that occur for each x

When ŧ(x) is estimated, heap height, say $h(x_h)$ is estimated by the difference between f(x) and ŧ(x) as specified in Equation 6. In other words, the height of a heap is the difference between the function value at the heap and the value of the trend through the f(x) for non-heap values.

Equation 6: $h(x_h)=f(x_h)-$ ŧ(x_h) where x_h is a heap

Given one does not know δ(x), using an approximation to it in applying Equations 2 to 4 is reasonable. As one sees from Equation 3, it is the shape of δ(x) that matters in computing proportions. A function roughly proportional to δ(x) will yield proportions similar to those calculated using δ(x). The function that is available and should be similar in shape to δ(x) is ŧ(x). However, if using ŧ(x) gives an approximation to δ(x), say $d_0(x)$, then it is reasonable to use $d_0(x)$ to get a better approximation of δ(x). Given that approximation, say $d_1(x)$, why not make another approximation using $d_1(x)$? In fact, one can think of going on to $d_2(x)$, $d_3(x)$, etc. until further iteration makes so little difference that more iteration is pointless. Equation 7 and 8 express using estimates from an iteration, n, in obtaining $d_{n+1}(x)$.

Equation 7: $P_{d,n}(x, [F,V,U])=d_n(x)/(\sum_{j=L}^{U} d_n(j))$

Equation 8: $R_{d,n+1}(x, [L,V,U])= P_{d,n}(x, [F,V,U])\,\eta([F,V,U])$
where
and η([F,V,U]) is the observed number in the heap related to [F,V,U]

A remaining matter regarding deheaping is dealing with prototypes that share heaps, heap to the same location. A prototype is designated as $H_{[L,V,U]}$.This is because, for example, for 0-5 heaping, the 5-width prototype [8,10,12] and the 11-width prototype [5,10,15] can both be used resulting in heaping on 10. In other words the size of the heap at 10 is $H_{[8,10,12]}$ plus $H_{[5,10,15]}$. Given some people may "round" to multiples of 5 near 20, 30 and even larger multiples of 10, shared heaping can occur at these values. In this research, only 5-heapers and 10-heapers sharing a heap is considered though heaping on multiples of 7 is allowed. Discussion pursues the matter of why prototypes such as [7,14,21] are not considered.

For multiples of 10, let h10 be a multiple of 10 and h(h10) be the estimated number of responses in the heap (i.e., h(h10)=f(h10)- ŧ(h10). For deheaping, a ratio, ρ_5 (h10), is needed that gives the proportion of h(h10) by 5-heapers. Since one does not know (h10), it must be estimated. For example, for 10 one has 5-heaps at 5 and 15 to use in approximating the proportion that 5-heapers are in the heap at 10. For the moose hunting data of Figure 2, computation of ρ_5(h10) must take into account that 7-related heaps occur at 7 and 14. Computations given in Appendix 2 result in ρ_5(10) =0.6 and ρ_5(20)=0.3. Unless someone can arrive at an exact computation method, a deheaping program must be provided with computed values of ρ_5(h10).

Finally, consider two prototypes such as [5,10,15] and [15,20,25]. Prototypes centered on even numbers overlap at boundaries of 15, 25, etc. and have a boundary a 5. You could say they are fuzzy at their boundaries. One can think of a person being relatively certain that the appropriate exact response, x, was between 10 and 20 but being more comfortable responding 10 or 20 than 15, the person is a 10-heaper. A logically consistent way of dealing with overlapping prototype boundaries is giving the boundary points a weight of half in distributing the number of responses in heaps back to x-values. This just acknowledges that the odds are 50/50 of heaping high or low (e.g., up to 20 or down to 10) for values other than 5. If a weight of a half is appropriate for other boundaries where responses are 5 from the heaping value, a weight of a half seems more reasonable for 5 than a weight of one. There simply is no research on this matter and assigning a weight of 0.5 or 1.0 makes little difference in estimation.

Implementation of Deheaping

A first consideration in deheaping a function is identifying its heaps. Figures 1 is useful in understanding why identifying heaps is done by the researcher. If you do not see heaps, deheaping is not necessary because nothing is there to influence estimates. However, if heaps are apparent, then use of a deheaping program for heaping caused by prototype use may be appropriate. Use of “may” is because a frequency function having heaps does not mean that heaping is resulting from the use of prototypes. Selling licences for say 4 days and for eight days can cause heaps at 4 and 8 because some purchasers will report using all days. Such heaping reflects administrative procedures thus prototype deheaping is inappropriate because responses reflect what was done. Therefore, a first step in deciding to apply prototype deheaping is identifying heaps that occur because of prototype use.

Deheaping being modeled is only for prototype use. Therefore, for now consider a f(x) with only prototype use heaps occuring. In Figure 1, heaps are seen at 5, 7, 10, 14, 15, 20, 25.and possibly at 21. Consider the high frequency at 30 is because 30 is either for 30 or more or is a limit on x (e.g., only 30 days of hunting a given species is allowed in a year). If it is a heap because 30 days or a total harvest of 30 is all that is allowed, it has a different implication for estimates of totals or means (e.g., mean days of hunting or total person days of hunting) than if large responses are pooled. This matter is pursued in the Discussion. Deheaping for the f(x) of Figure 1 is for responses less than 28. Dealing with heap at 28 with [25,28,31] would involve dealing with responses at 30. Dealing with 30 is not possible whether it is an upper limit or there is pooling since the authors have no way of estimating h(30).

Given a function has only prototype heaps except at a maximum value, material presented shows how to deheap. As specified in Equation 6, when heaps are identified and a trend below them exists and is estimated by regression, their location can be used in estimating the number of responses in heaps. The trend below the heaps provides a basis for starting to prorate the amounts in heaps to the x-values of prototypes that result in the heaps. Equations 7 and 8 algebraically define iteratively prorating responses in heaps back to where they came from based on the trend and then based on successive approximations to the deheaped function Therefore, as identified in Figure 2, programming the steps specified is how the deheaping program was implemented.

Using the Deheaping Program Implemented and Standard Deviations in Estimates

The deheaping program that is available operates as described above. It is available on the web (see Appendix 1) with documentation and examples. One is instructed on providing correct information to the program and running it in SAS© (? Reference). There are data and other material that allows one to confirm that the program is working. Though various adjustments can be made by changing program parameters, a user can proceed by providing information on where heaps are located and may modify information on the proportion of responses in heaps that are at multiples of 10 are 10-heapers.

As indicated above, Figure 2 gives general information on the flow of the deheaping program. The material above has focused on determining an estimate of what responses would be if there was no heaping. However, collection of data on days of hunting or harvest allows computing single numbers that are useful for managers and planners. If you know the total number of days, T, that N people spent hunting , say moose, mean days per person is u=T/N. Actually, if you are a resource manager, you may be interested in u and in T. Knowing u does not tell you the "load" being put on a resource. Knowing total numbers using a resource, N, does not give you a good measure of load. Knowing N and u over time can give you a perspective on how use is changing. The purpose of deheaping is to determine if data with heaping produces the same results as deheaped data. So, from the program one has non deheaped results (say u_n and T_n, Equation 3) and deheaped results (say u_d and T_d, Equation 4). How xf(x) should be dealt with for a pooled responses is addressed in the Discussion. Since N is not a value that should be determined by a deheaping program, it does not appear with subscripts but rather if the "N" implied by deheaping differs from N, the program makes a correction. In the context provided, a measure of the effect of heaping can be referred to a s bias and computed as in Equation 5. From Equation 5 you see that bias in the mean and total is the same in percent terms.

Equation 9: $T_n = \sum xf(x)$ over responses, over values of the frequency function

Equation 10: $T_d = \sum x\hat{d}_n(x)$ over values of n^{th} iteration estimate of the deheaped function

Equation 11: Bias Percent$=100(T_n-T_d)/T_d$

$=100(T_n/N-T_dN)/T_d/N$

$=100(u_n-u_d)/u_d$

Figure 4 about here

Figure 5:Example Results

Given information supplied to the program is correct, the program produces results like:

Bias percent is 2.8% with the estimated % of respondents heapers= 25.8%
and with observed total dayshunt=144243 while deheaped total dayshunt=140150 .
NOTE: Deheaped total dayshunt corrected for # of respondents deheaped #= 140275
Adjustment is because 19371 respondent are in the data.
This is a % difference of .09%=100*17/19371%.
Deheaping involves regression estimates giving deheaped # of respondents as 19354 .
Observed mean dayshunt is 7.45 while the deheaped mean is 7.25.
NOTE: Positive bias shows observed is too large by the % given.
NOTE: Bias is corrected for for observed number of responses NE to deheaped number.

Based on 20 randomization steps,
the mean bias for the frequency function is 2.66 with standard deviation of 0.076

The variety of information in Figure 3 is given to satisfy needs of both managers and researchers. A manager may just want to know that bias is 2.8% and know corrected and uncorrected estimates of means or totals (e.g., of mean days hunting moose and total person days hunting moose). This may be low enough that concern only arises when a change in data collection method results in a change in bias. A small change in bias can matter as one can get the false impression of mean or total days hunting starting to change at a rapid rate. Knowing that about 25% of responses is in heaps can be of value to a researcher who wishes to compare the amount of heaping that occurs with different survey methods (e.g., with a study such as Beaman., Vaske and Miller 2005b). Knowing that deheaping results in a slightly different total number of respondents may also interest a researcher. The difference in total results from the trend established by regression not having the same total as the observed responses that are not for heaps. Knowing that the totals are within a certain percent is evidence that the trend fit was good or poor.

An important matter if one is using estimates is having standard deviation of the estimates. The estimation program provides a vehicle for deriving standard deviations. The standard deviations in bias and some other estimates are computed using randomization (obtain program documentation as instructed in the Appendix). By introducing the random variation that is occurring in observed values into values of the f(x) for which estimates are made, one finds out how much random variation can be influencing estimated values. Any f(x) being deheaped that is based on N observations has $p(x)=f(x)/\Sigma f(x)$ for the estimated probability of a response x. Given p(x) is relatively small, the number of responses of x to expect in N trials is approximately Poisson distributed. By taking f(x) to be an estimate of μ, one also has an estimate of σ^2 since for the Poisson, $\mu= \sigma^2$. By creating multiple random "influenced" versions of f(x) and making estimates for these, one gets estimates of means and totals that would occur given variation likely occurring in multiple versions of f(x). Standard deviations in estimates can be computed from the, say 20, alternative values of estimates produced by runs on 20 "randomizations".

If one has an estimated of bias of 2% and its standard deviation is 4%, there is not much point in doing anything with the estimate. Estimates from f(x) are as good as you are going to get. However, above you that for an estimate of 2.55% the standard deviation is about .1. You have evidence that using f(x) produces estimates that are high. However, above it was made clear that a function c(x) and proportions of 5 to 10 heapers have been used in estimation. These "parameters" of the estimation are not exact. Varying these can influence estimates. The influence of such change is discussed below under the heading Sensitivity.

Discussion

This paper builds on Beaman, Vaske and Grenier (1998). By using regression to determine a trend showing where heaps begin, the paper makes more systematic use of information than the earlier approach. Another advance over the earlier work is recognizing that the trend function that is used in determining the height of heaps is only a first approximation to the shape of the deheaped function that that should be used in distributing response in heaps appropriately. In other words, the need to iteratively move from the trend function to successive approximations of the deheaped function is a contribution. Also, recognizing the need to estimate the relative numbers of 5 and 10-heapers at heaps that are a multiple of 10 is impotant. Addressing the tendency to avoid certain responses (i.e., re 9, 11 and 13 in Figure 2) and allowing for this is another contribution. Finally, producing standard deviations in estimates by recognizing that having standard deviations in estimates can be achieve by randomization and repeated estimation is a contribution.

One may wonder why this article is relevant when one has mathematically sophisticated deheaping procedures available. Camarda, Eilers and Gampe (2008) is an example of sophisticated modelling of general patterns of digit preference. The difference between the mathematical approach and this paper is pursuing smoothing based how people arrive at responses. A mathematical distribution of heaps need not, and thus may not, reflect how people behave. Given mathematical deheaping need not reflect behavior, logic suggests using the deheaping approach presented here and addressing behavorial issues on which research is still needed.

Because this paper is about implementing deheaping when heaps result from prototype use, success is shown by results provided in Table 5. Given that a frequency function to be deheaped only has heaps at multiples of 5 or 7, the program produced and freely available (see Appendix 1), achieves deheaping and estimates as described. Therefore, the limitations and needs for enhancement of the deheaping program are pursued under the heading "The Need for Further Research and Program Enhancement."

The Need for Further Research and Program Enhancement

A matter for further research is developing a program that does not depend on a trend, t(x), as a critical step in estimating the distribution of responses that would occur if there were no heaping. When one has heaps with few non heaped responses, the program developed is not appropriate since estimation depends on developing a trend function base on most y-values being greater than zero. The program is not useful for much expenditure data since, in the authors' experience expenditure data tends to be a collection of heaps. Also, the program is for deheaping when respondents have equal probability of replacing any responses in an interval by a central value. This has been referred to as unbiased responding and has been taken to be an attribute of prototypes being used. For wildlife or other data in which people respond even with 50, it may be clear that a person is more likely to round down from, say 60, than to round up from 40. This is not unbiased use of a prototype and the program is not appropriate.

Mention of larger responses relates to a limitation on the deheaping procedure relating to the largest response (i.e., maximum x-value) being a pooled response such as “30 or more”. If 232 respondents reply 30 because 30 days was the maximum days or harvest allowed, then 232*30 is the number of person days to use in determining total person-days and mean days per person. However, if 30 refers to “30 or more” then 232*30 is an underestimate, given most actual values if reported would be greater than 30. As one sees in Figure , program output provides information on how a number like 233*30 (i.e., max(x-value)*number giving that response) relates to total person-days.

The matter of estimates made being sensitive to functions used in estimation and prototype use being unbiased has been mentioned. A program enhancement that is being pursued is solving for best values of ρ_5(h10). By altering values and assessing “goodness” of deheaping, the theory is that best values of ρ_5(h10) can be determined. This research presents challenges since defining goodness of fit cannot be how an estimate curve relates to observations since the goal is producing a function from heaped data that has no heaps. As for a prototype [L,V,U] being such that V=(L+H)/2, this may not be a good assumption as responses become large (e.g., 50 or greater). More needs to be known about prototypes used in giving large responses.

While this research makes advances over Beaman, Vaske and Grenier (1998) research on may matters will allow improvements. In particular, formation on 5 and 10-heapers sharing responses in heaps at multiples of 10 is particularly weak. Therefore, research such as using computer assisted data collection in asking about the way responses at heap-values are arrived at is important. Such computer assisted research can throw light on preference for responses that do not end in 1, 3, 7 and 9, with exceptions like the disposition to 7 when thinking in weeks. In fact, in a project where a computer introduces questions based on x-values, one could press respondents on the “best” value they can arrive at for responses of 40 or larger and thus provide new insights into prototype use. Given that responses of 1 or 2 are probably considered accurate by respondents and modelling all of a sudden has uncertainty associated with three or greater, probing on certainty about small responses could also be fruitful.

Mathematical (not behavior based) deheaping programs are an alternative to the program reported on here. Camarda, Eilers and Gampe (2008) have provided R-code so what their program produces can be compared with the deheaping program of this research. This comparison is being pursued to understand why differences in estimates occur. One avenue is modifying R code so mathematical and behavior based estimates tend to agree. Given “R” is free, having a simple approach to estimation using the “R” statistical package that gives about the same results as using SAS@ would be advantageous for those who do not have access to SAS@.

Conclusion

As stated above, this research improves on earlier work by introducing important considerations not pursued in the past deheaping research. However, the computer program implementing concepts into estimation only allows for deheaping to occur when heaps are due to prototype use and are at multiples of 5 and 7. Even then dealing with heap sharing (e.g., 5 and 10 heapers both heaping to multiples of 10) requires program users to specify proportions such as the percent of responses in the heap at 10 that come from 5-heapers). In that regard, a better understanding of prototype use, particularly shared heaping, is needed. Understanding of concepts is key to improved quantitative application of the concepts. Still, the program developed and made available on the web is adequate for addressing the effect of heaping on means and totals of data on such matters as days spent (e.g., days hunting or fishing) or harvest (e.g., ducks shot or fish caught). Not only providing an estimate of bias but an estimate of the standard deviation in that bias makes it possible to recognize if a bias estimate is statistically meaningful. In other words, the work advances research and management use of information.

References

Beaman, J.G.; Vaske, J. ; Grenier, M. 1998. "A prototype model for estimating and correcting bias in digit preference / number preference" Tourism Analysis, 2(2):77-90.

Beaman, J. G.; Vaske, J.; & Huan, T. C. 2003. "The Use Of Multiples In Anglers' Recall Of Participation And Harvest Estimates: Some Results And Implications". Leisure Sciences, 25:399-410.

Carlo G. Camarda, Paul H.C. Eilers and Jutta Gampe (2008) "Modelling general patterns of digit preference" Statistical Modelling 2008; 8(4): 385--401

Hultsman, W. Z., Hultsman, J. T., & Black, D. R. (1989). "Response peaks" as a component of measurement error: Assessment implications for self-reported data in leisure research. *Journal of Leisure Research*, *21*(4), 310–315.

SAS Institute Inc., (1989) SAS/STAT© User's Guide, Version 6, 4th Ed., Vol. 2, Cary, N.C.

Vaske, J. J., Beaman, J., Manfredo, M. J., Covey, D., & Knox, R. (1996). Response strategy, recall frame, and digit preference in self-reports of angling participation. *Human Dimensions of Wildlife, 1*, 54–68.

Vaske, J. , Beaman, J. (2006). "Lessons Learned In Detecting And Correcting Response Heaping:Conceptual, Methodological And Empirical Observations". Human Dimensions of Wildlife 11:285-296.

Yates, F. (1934) "Contingency tables involving small number and the χ^2 test" Supplement to the Journal of the Royal Statistical Society 1(2): 217–235.

Figure 1:

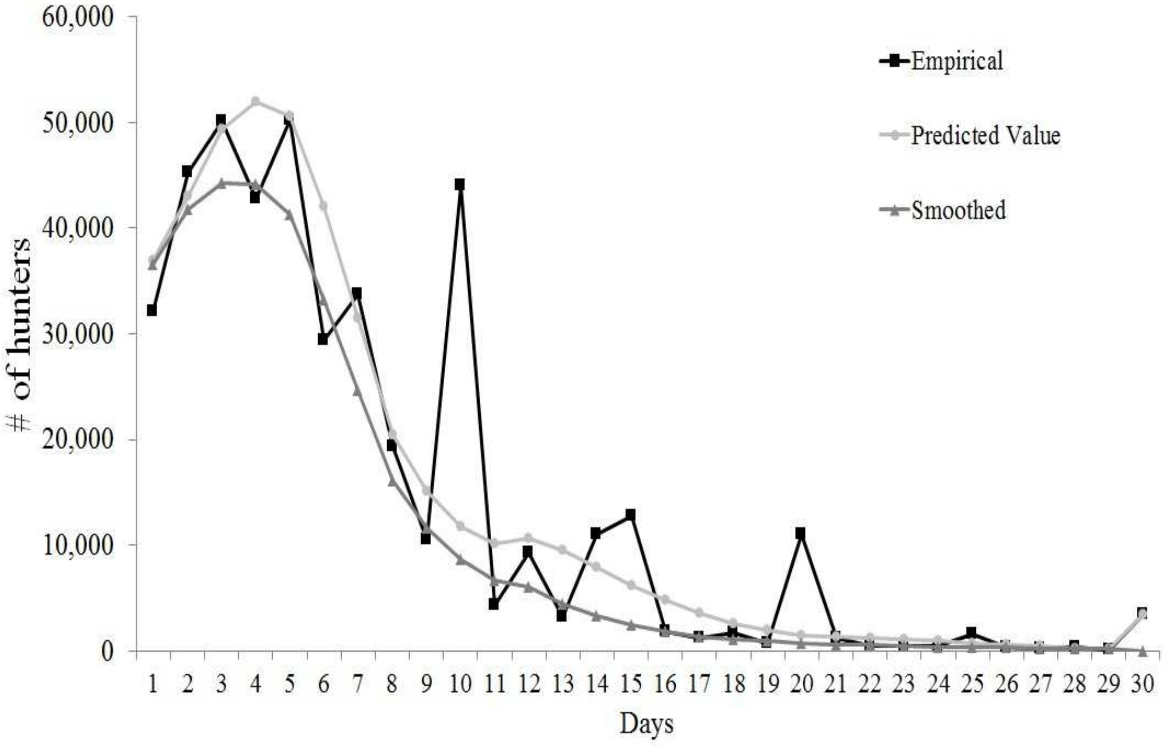
60,000
50,000
40,000
30,000
20,000
10,000
0
# of hunters
Empirical
Predicted Value
Smoothed
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Days

Figure 2: Program steps associated with getting the program running

1. Provide information for the program to access and store data

2. Read a frequency function and store values needed for deheaping

3. Process heap location information and store for use in deheaping

4. Examine the data to create code for regressions for estimating a smooth curve showing where heaps start and run the regressions.

5. Using regression results get heap heights.

6. Remove an estimate of responses made with certainty from the “trend below the heaps” to get a “trend in uncertain responses”.

7. Distribute heaps using the “trend in uncertain responses” and smooth the distributed results for adding them back to the “trend below the heaps” (the total count at heaps at particular “x” values is broken up and distributed based on proportions derived from the “trend in uncertain responses”).

8. Proceed with 5 iterations in which estimates of the deheaped function are improved by using a the previous estimate of the deheaped function, with the estimate of certain responses removed) to get proportions to distribute the total count at heaps more appropriately.

9. Output the estimated deheaped frequency function and bias related statistics such as the means and totals (e.g., mean days hunting and total person days hunting) computed from observations and the deheaped frequency function. Also, compute differences in these as a percent (i.e., estimate bias). See Figure 4 for actual results.

10. Introduce appropriate (Poisson) random variation into the observed frequency function and make estimates for these “randomizations” (e.g., 20 times) to get standard deviations in estimates (e.g., in bias in average days hunting or total person days of hunting and output the results—see Figure 4).