Journal of Data Science **3**(2005), 439-448

Sampling Random Variables: A Paradigm Shift for Opinion Polling

Gordon G. Bechtel
University of Florida and Florida Research Institute

Abstract: Conventional sampling in biostatistics and economics posits an individual in a fixed observable state (e.g., diseased or not, poor or not, etc.). Social, market, and opinion research, however, require a cognitive sampling theory which recognizes that a respondent has a choice between two options (e.g., yes versus no). This new theory posits the survey respondent as a personal probability. Once the sample is drawn, a series of independent non-identical Bernoulli trials are carried out. The outcome of each trial is a momentary binary choice governed by this unobserved probability. Liapunov's extended central limit theorem (Lehmann, 1999) and the Horvitz-Thompson (1952) theorem are then brought to bear on sampling unobservables, in contrast to sampling observations. This formulation reaffirms the usefulness of a weighted sample proportion, which is now seen to estimate a different target parameter than that of conventional design-based sampling theory.

Key words: Independent non-identical (i.n.d.) distributions, personal probability, sampling unobservables, self weighting, target parameter, variance estimation, weighted Bernoulli trials.

1. Introduction

Estimating a population proportion is commonplace in psychological assessment, experimentation, and opinion surveys. In this vast area of research and application the target parameter is invariably posed as a *single* probability that governs a sequence of binary respondent outcomes such as "right vs. wrong", "agree vs. disagree", etc. The classical central limit theorem is then used to support the normality of the sample proportion.

Oddly, the above status quo in social research and policy application ignores a well known fact; namely, the existence of broad individual differences on almost any psychological variable imaginable. From a statistical point of view data collectors and analysts have been content with the central limit theorem of Laplace in 1810 rather than progressing to that of Liapunov in 1901. (These sources are referenced by Lehmann 1999, pp.600-601). When applied to opinion polling, the earlier theorem assumes a population of individuals, each having an *a priori* agreement or disagreement with a statement not yet heard. Thus the presentation of this statement to a sample of individuals from this population is tantamount to the selection of red and white balls from an urn, with the goal of estimating the proportion of red. That is, successive individual-by-individual response solicitations are regarded as independent identically distributed (i.i.d.) Bernoulli trials, each with a common probability that is the population proportion.

The present paper argues for the application of the more realistic Liapunov central limit theorem. This relaxes the status quo to independent *non-identically* distributed (i.n.d.) Bernoulli trials, each with an individual-specific (case) weight and response probability. First, a sample of size n is drawn without replacement from a population of N individuals. The sample design determines the inclusion probability for each individual in the population. Next, conditioning on the selected sample, the data collection consists of n solicited i.n.d. Bernoulli responses. Rather than revealing a predetermined individual state, each Bernoulli trial generates a momentary response (a one or zero) driven by an individual's unobserved probability. Given this alternative representation of the survey respondent, the Liapunov central limit theorem, and the Horvitz-Thompson (1952) theorem, are then invoked to estimate the mean of the population of personal probabilities.

Section 2 lays out a triangular array of i.n.d. random variables and its central limit theorem. Section 3 defines a weighted Bernoulli variate and inserts it into this array. Section 4 restricts and interprets this formulation to design-based sampling (without replacement) from a finite population of random variables. This gives the conditional and unconditional expectations of the (approximately normal) sample mean of n weighted Bernoulli variates, along with its conditional variance. Section 5 treats the important and frequently used case of self weighted, or *epsem* (equal probability of selection method), samples. In this case the classical estimate of the standard error of the sample proportion is advocated as reasonable rather than the severe underestimate it is widely believed to be. Finally, Section 6 gives some concluding remarks relevant to this new world of survey sampling.

2. The Central Limit Theorem for I.N.D. Random Variables

We begin with a sequence of sets of random variables of size n in the triangular array

$$Y_{ni}, \quad i = 1, 2, \ldots, n$$

where $n \to \infty$. The following notation will be used to describe $Y_{n1}, Y_{n2}, \ldots, Y_{nn}$:

$$E(Y_{ni}) \quad = \quad \mu_{ni}$$

$$
\begin{aligned}
Var(Y_{ni}) &= \sigma_{ni}^2 < \infty \\
\bar{\mu}_n &= \frac{1}{n}(\mu_{n1} + \mu_{n2} + \cdots + \mu_{nn}) \\
s_n^2 &= \sigma_{n1}^2 + \sigma_{n2}^2 + \cdots + \sigma_{nn}^2 \\
\bar{Y}_n &= \frac{1}{n}(Y_{n1} + Y_{n2} + \cdots + Y_{nn}).
\end{aligned}
$$

Using this notation, a convenient lemma to Liapunov's theorem (cf. Lehmann, 1999, pp. 97-102, 571-573) may then be invoked:

Lemma 2.2.1 Let Y_{ni} be a typical element of a triangular array of independent random variables with $E(Y_{ni}) = \mu_{ni}$ and $Var(Y_{ni}) = \sigma_{ni}^2$. Then

$$\frac{\bar{Y}_n - \bar{\mu}_n}{\sqrt{Var(\bar{Y}_n)}} \to N(0,1) \quad \text{as } n \to \infty \tag{2.1}$$

provided there exists a constant A such that

$$|Y_{ni}| \leq A \ \text{ for all } \ i \text{ and } n \tag{2.2}$$

and

$$s_n^2 \to \infty \tag{2.3}$$

Conditions (2.2) and (2.3), which are jointly sufficient for (2.1), imply the sufficient condition used by Liapunov to establish (2.1) (Lehmann,1999, pp. 98, 101). The more restrictive conditions (2.2) and (2.3) are used here because they are easily satisfied by the weighted Bernoulli variates defined next.

3. I. N. D. Bernoulli Trials

For individual i let X_{ni} be a Bernoulli variate taking the values 1 or 0 with probabilities p_{ni} and $1 - p_{ni}$. Now define i's weighted Bernoulli variate as

$$Y_{ni} := w_{ni} X_{ni}, \tag{3.1}$$

where $w_{ni} > 0$ and sum of the weights is $\sum_i w_{ni} = n$. Then

$$
\begin{aligned}
E(Y_{ni}) &= \mu_{ni} = w_{ni} p_{ni} \\
Var(Y_{ni}) &= \sigma_{ni}^2 = w_{ni}^2 p_{ni}(1 - p_{ni}) < \infty
\end{aligned}
$$

The value taken by Y_{ni} is uniformly bounded, satisfying (2.2). Moreover, in this binary situation

$$s_n^2 = \sum_i w_{ni}^2 p_{ni}(1 - p_{ni}) \tag{3.2}$$

satisfies (2.3) when the p_{ni} are bounded away from zero and one. That is, if their exists a constant $a > 0$ such that

$$a < p_{ni} < 1 - a \quad \text{for all } i \text{ and } n,$$

then $1 - p_{ni} > a$ and $p_{ni}(1 - p_{ni}) > a^2$. Also, the weights w_{ni} vary about one, and there exists a constant $b > 0$ such that

$$w_{ni} > b \quad \text{for all } i \text{ and } n.$$

Hence, $w_{ni}^2 > b^2$,

$$w_{ni}^2 p_{ni}(1 - p_{ni}) > b^2 a^2,$$

and

$$s_n^2 = nb^2 a^2 \to \infty \text{ as } n \to \infty,$$

verifying (2.3) (cf. Lehmann, 1999, p. 99).

4. Sampling Random Variables

4.1 A finite triangular array

With conditions (2.2) and (2.3) satisfied, the limiting distribution (2.1) is now established for the n weighted Bernoulli variates in (3.1). These i.n.d. random variables $Y_{n1} \ldots, Y_{ni}, \ldots, Y_{nn}$ in lemma 2.2.1 are now interpreted as arising from *a particular* sample from a finite population. Therefore we must now restrict the array in Section 2 to a finite set of random variables and interpret these as a sequence of samples of size $n = 1, \ldots, N$. This setup is depicted by the array

$$Y_{ni}, \quad i = 1, 2, \ldots, n; n = 1, 2, \ldots N$$

where the n-th member of this sequence is a sample of n random variables drawn (without replacement) from the N-th member, which is a population of N random variables. It is reiterated that respondent i in this nth sample is represented here as a *random variable* Y_{ni} upon which an observation is to be realized during later response solicitation.

Curtailing the infinite sequence in Section 2 to the finite population $Y_{N1}, \ldots$, Y_{Ni}, $\ldots, Y_{NN}$ weakens the *asymptotic* normality of $\bar{Y}_n$ to its *approximate* normality. Also, any simple or complex sampling design determines a probability $\pi_{ni} > 0$ that individual i is included in the sample. For example, if the design is self weighting this inclusion probability is $n/N$for each individual in the population. (See Section 5.) Finally, in this special *sampling* case of lemma 2.2.1, the weights of the i.n.d. Bernoulli variates take the form

$$w_{ni} = \frac{n}{N\pi_{ni}}, \tag{4.1}$$

where again $w_{ni} > 0$ and $\sum_{i \in S} w_{ni} = n$.

4.2 The conditional and unconditional expectations of $\bar{Y}_n$

We now condition on the i.n.d.random variables $Y_{n1}, \ldots, Y_{ni}, \ldots, Y_{nn}$ actually drawn, observing that their (approximately) normal mean,

$$\bar{Y}_n = \frac{1}{n} \sum_{i \in S} Y_{ni} = \frac{1}{N} \sum_{i \in S} \frac{X_{ni}}{\pi_{ni}}, \tag{4.2}$$

has the sample specific expectation

$$E(\bar{Y}_n) = \frac{1}{N} \sum_{i \in S} \frac{p_{ni}}{\pi_{ni}} =: p_n \tag{4.3}$$

The sample sum in (4.3) is, by the Horvitz-Thompson (1952) theorem, an unbiased estimate of the total of the N response propensities p_{Ni} in the population. This is stated in the following lemma:

Lemma 4.2.1 Let $T_N = p_{N1} + \cdots + p_{NN}$ be the sum of theN unobserved probabilities in the population. Then

$$T_N = E\left(\sum_{i \in S} \frac{p_{ni}}{\pi_{ni}} \right). \tag{4.4}$$

The Horvitz-Thompson estimator in the parentheses in (4.4) is an unbiased estimate of the population total for an arbitrary sampling design (Thompson 1997, pp.12-15; Lohr 1999, pp.196-199, 204-210). Thus lemma 4.2.1 implies that the sample-specific expectation p_n under lemma 2.2.1 itself has the expectation

$$E(p_n) = \frac{T_N}{N} = E\{E(\bar{Y}_n)\}, \tag{4.5}$$

over all samples of size n. This latter expectation is the population mean of the N *unobservable* probabilities p_{Ni} . This population mean is also the unconditional expectation of the *observed* sample mean $\bar{Y}_n$ in lemma 2.2.1.

4.3 A population census

In a census $n = N$, and

$$\bar{Y}_N = \frac{1}{N}(Y_{N1} + \cdots + Y_{NN}).$$

Substituting the inclusion probability of one for π_{ni} in (4.3)

$$E(\bar{Y}_N) = P_N.$$

Interestingly, the target parameter P_N, which is the mean of the population of N personal probabilities, is not realized but only *expected* in a census. That is, in a census the mean $\bar{Y}_N$ is still a random variable because each individual response Y_{Ni} is stochastic, taking the values 0 with probability $1 - p_{Ni}$ and the value 1 with probability p_{Ni}, for $i = 1, \ldots, N$. With $w_{Ni} = 1$ the variance of this census mean is easily seen to be

$$Var(\bar{Y}_N) = \frac{1}{N^2} \sum_{i \in P} p_{Ni}(1 - p_{Ni}), \tag{4.6}$$

where the summation is over $i = 1, \ldots, N$ for the entire population. The miniscule variance in (4.6) shows that the census $\bar{Y}_N$ is a *random variable* that is distributed tightly around the target parameter P_N . In contrast, in standard design-based surveys (Lohr, 1999) the census mean is the *fixed* proportion of 1's (versus 0's) in the population.

4.4 The conditional variance of $\bar{Y}_n$

The variance of the sample mean $\bar{Y}_n$ in (4.2), which is conditioned on the sample, may be alternatively expressed as

$$\begin{aligned} Var(\bar{Y}_n) &= n^{-2} s_n^2 & (4.7) \\ &= N^{-2} \sum_{i \in S} \frac{p_{ni}(1 - p_{ni})}{\pi_{ni}^2} & (4.8) \end{aligned}$$

where s_n^2 is given in (3.2). Writing (4.7) as

$$Var(\bar{Y}_n) = n^{-1} \frac{s_n^2}{n},$$

the variance of the sample mean is seen to be the mean of the n variances divided by the sample size n. This is a generalization of the classic special case, where the variance of the sample mean is the (single) population variance divided by the sample size.

Finally, the Horvitz-Thompson (1952) theorem can also be applied to the unobserved individual variances $p_{ni}(1 - p_{ni})$ in (4.8):

Lemma 4.4.1 Let $V_N = p_{N1}(1 - p_{N1}) + \cdots + p_{NN}(1 - p_{NN})$ be the sum of the N unobserved variances in the population. Then

$$Var(\bar{Y}_N) = \frac{1}{N^2} V_N = \frac{1}{N^2} E \left\{ \sum_{i \in S} \frac{p_{ni}(1 - p_{ni})}{\pi_{ni}} \right\}. \tag{4.9}$$

Writing the expectation of the variance of the sample mean in (4.8) as

$$E(Var(\bar{Y}_n)) = \frac{1}{N^2} E\left\{\sum_{i \in S} \frac{p_{ni}(1-p_{ni})}{\pi_{ni}} \cdot \frac{1}{\pi_{ni}}\right\} \tag{4.10}$$

it becomes evident that

$$E(Var(\bar{Y}_n)) > Var(\bar{Y}_N)$$

due to the multiplier $\pi_{ni}^{-1} > 1$ in (4.10). In the important case of self weighting in Section 5, $\pi_{ni}^{-1} = N/n$ and therefore

$$E(Var(\bar{Y}_n)) = \frac{N}{n} Var(\bar{Y}_N) \tag{4.11}$$

Equation (4.11) shows that the conditional variance of the mean $\bar{Y}_n$ is a different order of magnitude than the (miniscule) variance of the census mean $\bar{Y}_N$. A particular estimate of $Var(\bar{Y}_n)$ is suggested for the case of self weighting.

5. Self Weighting

Complex surveys commonly use stratified multi-stage sampling with all units selected with probability proportional to size except at the final stage. In this last stage a fixed number of individuals are drawn from the last unit (e.g., voting district) by simple random sampling without replacement. This sampling design is self weighting in the sense that each individual in the population has the same probability n/N of being included in the sample. (Skinner, Holt, and Smith, 1989, pp.16, 40; Thompson, 1997, pp.12-15). Other types of *epsem* designs are used in random-digit-dialing telephone surveys in marketing research. Self weighting also occurs in simple surveys, where n individuals are drawn directly from a population of size N by simple random sampling without replacement.

5.1 The conditional epsem variance of $\bar{Y}_n$

Substituting n/N for π_{ni} in (4.1) gives

$$w_{ni} = n(Nn/N)^{-1} = 1 \qquad \text{for } i = 1, 2, \ldots, n$$

Replacing w_{ni}, in turn, by one in (3.2) and (4.7) gives

$$Var(\bar{Y}_n) = n^{-2} \sum_{i \in S} p_{ni}(1-p_{ni}). \tag{5.1}$$

Formula (5.1) is also found by substituting n/N for π_{ni} in (4.8). It is then easily shown that

$$\sum_{i \in S} p_{ni}(1 - p_{ni}) = np_n(1 - p_n) - \sum_{i \in S}(p_{ni} - p_n)^2 \tag{5.2}$$

Replacing π_{ni} by n/N in (4.3), reveals that p_n in (5.2) has the structure

$$p_n = \frac{1}{n}(p_{n1} + p_{n2} + \cdots + p_{nn}),$$

which is the mean of the n individual probabilities controlling the sampled Bernoulli variates $Y_{n1}, \ldots, Y_{ni}, \ldots, Y_{nn}$. Finally, dividing both sides of (5.2) by n^2 gives

$$Var(\bar{Y}_n) = \frac{p_n(1 - p_n)}{n} - \frac{1}{n^2}\sum_{i \in S}(p_{ni} - p_n)^2, \tag{5.3}$$

showing that the conditional variance of $\bar{Y}_n$ increases as $p_{n1}, \ldots, p_{nn}$ become more homogeneous, maximizing when these probabilities are all equal. Therefore, $p_n(1 - p_n)/n$ is an upper bound for $Var(\bar{Y}_n)$ in (5.3).

5.2 Inferences from $\bar{Y}_n$ to p_n and P_N

Equation (5.3) suggests the classical statistic $\bar{Y}_n(1-\bar{Y}_n)/n$ as an *over*estimate of the conditional (sample dependent) variance in the self weighted case. Moreover, this conservative variance estimate holds for *all* sample sizes up to and including the population size N. Thus, even in a census, where

$$\bar{Y}_N = N^{-1}\sum_{i \in P} Y_{Ni}$$

the statistic

$$\frac{\bar{Y}_N(1 - \bar{Y}_N)}{N}$$

is an *over*estimate of $Var(\bar{Y}_N)$ in (4.9).

Over-estimating $Var(\bar{Y}_n)$ in (5.3) as $\bar{Y}_n(1-\bar{Y}_n)/n$ sets up the *very* conservative confidence interval

$$\bar{Y}_n \pm 1.96\sqrt{\frac{\bar{Y}_n(1 - \bar{Y}_n)}{n}}, \tag{5.4}$$

which is greater than 95% for covering p_n. This interval is less conservative for covering P_N, which is generally more distant from $\bar{Y}_n$ than p_n .

Finally, it is well known that $\bar{Y}_n(1 - \bar{Y}_n)/n$ severely *under*estimates the variance of the sample proportion in conventional complex sampling, where each fixed-state respondent is represented by a 0 or 1. In this standard design-based situation the true variance of the proportion is inflated by the homogeneous clustering of 0's and 1's in the population. In contrast, the present construction represents the respondent by a Bernoulli trial that is driven by an unobserved personal probability. In this alternative, more realistic representation of the respondent, the statistic $\bar{Y}_n(1-\bar{Y}_n)/n$ is a reasonable estimate of the unconditional variance of $\bar{Y}_n$ over all samples of size n.

6. Discussion

The present paper advocates the design-based sampling of random Bernoulli variates versus the conventional design-based sampling of 1's and 0's. Both procedures produce an observed sample of 1's and 0's and a sample proportion, but their generating processes are very different. A sample of random variables, with subsequent Bernoulli trials, gives a sample proportion that estimates the mean of a population of proportions. In contrast, numerical samples give a sample proportion that estimates the mean of a population of 1's and 0's. Although these two sample proportions are identical, their variances and target parameters are quite distinct.

In the case of Bernoulli variate sampling, individual differences are treated in two senses. First, they are regarded as varying response dispositions governed by individual-specific probabilities p_{ni} . Second, these dispositions are attended by individual-specific weights w_{ni} that are differential representations of the individuals in a population. In this setup individual i's personal probability takes a value in the open interval (0,1), in contrast to i being (extremely) represented by either zero or one. Hence, the sampling is from a population of *unobservable* probabilities rather than a population of *observable* zeros and ones. The latter convention is appropriate in biomedical and economic research, where individual i is in a fixed and noticable state, such as diseased or not, poor or not, etc. In social, market, and opinion research, however, an individual has a choice of responding one way or the other. In the present formulation, this choice is under the control of a personal response disposition p_{ni} that is activated upon stimulus presentation. The response observed is still a zero or one but now these two values are taken by a random Bernoulli variate at the individual level.

An important strength of the present approach is that the individual propensity p_{ni} remains unobserved, allowing us to side step its estimation by complex numerical iterations or lengthy experimental replications. For example, item response theory requires computationally intensive methods to estimate distinct individual probabilities for saying “yes” to a survey question. On the other hand, signal detection theory uses arduous replications to estimate subject-specific probabilities for saying “yes” that a tone is present amid noise. Calculations such as these may be necessary for individual evaluation in psychology and education, but the Liapunov and Horvitz-Thompson theorems allow us to circumvent them for group assessment at the population level.

Finally, equation (5.3) provides a conventional estimate of $Var(\bar{Y}_n)$ in the context of the sampling theory developed in Sections 2 through 5. In the case of self weighting the homogeneity of the p_{ni} reduces the second term in (5.3), increasing $Var(\bar{Y}_n)$. Thus the estimate of this conditional variance, suggested in Section 5.2, is very conservative, providing a reasonable estimate of the unconditional variance of $\bar{Y}_n$. In the unweighted case this is a reassuring property for an ordinary sample mean because the *over*estimate $\bar{Y}_n(1 - \bar{Y}_n)/n$ of $\bar{Y}_n$'s conditional variance holds for all sample sizes up to and including the population size N. In contrast, this classical statistic severely *under*estimates the variance of the sample mean in standard complex sampling from populations with homogeneous clusters of zeros and ones. Thus, replacing a respondent's *spurious* zero or one by a Bernoulli trial driven by his (or her) personal probability provides a fresh look at the important issue of variance estimation in opinion surveys.

Acknowledgement

This work has been supported by the University of Florida's Warrington College of Business Administration.

References

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663-685.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory.* Springer.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* Duxbury Press.

Skinner, C. J., Holt D. and Smith, T. M. F. (Eds) (1989). *Analysis of Complex Surveys.* Wiley.

Thompson, M. E. (1997). *Theory of Sample Surveys.* Chapman and Hall.

Received February 29, 2004; accepted August 3, 2004.

Gordon G. Bechtel
University of Florida and Florida Research Institute
P.O. Box 117155
Gainesville, Florida 32611-7155
USA
bechtel@nersp.nerdc.ufl.edu